Erasmus Mundus Joint Master on Data Engineering and Artificial Intelligence (DEAI) Course Programme

Contents

Chapter 1

Introduction

Vision The DEAI programme aims at becoming a world-wide reference in the development of knowledge, talent and leadership in Data Engineering and Artificial Intelligence to cope with the lack of productivity in complex and large data-driven AI projects. Our vision is that the systematisation and operationalisation of cutting-edge AI projects is not possible without rigourous DE. Therefore, AI initiatives must rely on DE in order to boost Europe's competitiveness in front of the US and China in the AI field.

Mission We bring together world-leading institutions with a long experience in both research and teaching collaboration and define a concrete and realistic common mission statement as follows:

- Educate a large number of highly-qualified and innovative professionals in both DE and AI. This hybrid profile is in great demand in Europe and the world,
- Define and implement a curriculum that provides a holistic view of the field, meeting the innovation aspects required by industry and large-size AI projects who rely on highly complex data value chains,
- Promote the expertise of European HEIs on the field and attract excellent local and foreign students, with the final aim to boost Europe's expertise and close the current gap in AI with US and China,
- Develop and active community on the conjunction of DE and AI, facilitating the exchange of information and experiences in order to foster the knowledge triangle between education, research and innovation.

Chapter 2

ULB — Data Engineering Foundations (1st semester)

2.1 Detailed Description of Curriculum

This semester introduces essential topics in BI, covering data warehouses, data mining, and business processes, as well as essential aspects of data management, covering traditional relational technology and new emerging paradigms. It is composed of the following courses.

University: Université Libre de Bruxelles (ULB)
Department: École polytechnique de Bruxelles
Course ID: ADB (INFO-H-415)
Course name: Advanced Databases
Name and email address of the instructors: Esteban Zimányi (ezimanyi@ulb.ac.be)
Web page of the course: http://cs.ulb.ac.be/public/teaching/infoh415
Semester: 1
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 24h.
- Exercises: 24h.
- Projects: 12h.

Goals:
Today, databases are moving away from typical management applications, and address new application areas. For this, databases must consider (1) recent developments in computer technology, as the object paradigm and distribution, and (2) management of new data types such as spatial or temporal data. This course introduces the concepts and techniques of some innovative database applications

Learning outcomes:
At the end of the course students are able to

- Understand various different technologies related to database management system
- Understand when to use these technologies according to the requirements of particular applications
- Understand different alternative approaches proposed by extant database management systems for each of these technologies
- Understand the optimization issues related to particular implementation of these technologies in extant database management systems.

Readings and text books:

- R.T. Snodgrass. *Developing Time-Oriented Database Applications in SQL*, Morgan Kaufmann, 2000
- Jim Melton, Alan R. Simon. *SQL: 1999 - Understanding Relational Language Components*, Morgan Kaufmann, 2001
- Jim Melton. *Advanced SQL: 1999 - Understanding Object-Relational and Other Advanced Features*, Morgan Kaufmann, 2002
- Shashi Shekhar, Sanjay Chawla. *Spatial Databases: A Tour*, Prentice Hall, 2003.

Prerequisites:

- Knowledge of the basic principles of database management, in particular SQL

Table of contents:

- Active Databases
  Taxonomy of concepts. Applications of active databases: integrity maintenance, derived data, replication. Design of active databases: termination, confluence, determinism, modularisation.
- Temporal Databases
  Temporal data and applications. Time ontology. Conceptual modeling of temporal aspects. Manipulation of temporal data with standard SQL. New temporal extensions in SQL 2011.
- Object-Oriented and Object-Relational Databases
  Object-oriented model. Object persistance. ODMG standard: Object Definition Language and Object Query Language. .NET Language-Integrated Query: Linq.
  Object-relational model. Built-in constructed types. User-defined types. Typed tables. Type and table hierarchies. SQL standard and Oracle implementation.
- Spatial Databases
  Application domains of Geographical Information Systems (GIS), common GIS data types and analysis. Conceptual data models for spatial databases. Logical data models for spatial databases: rastor model (map algebra), vector model (OGIS/SQL1999). Physical data models for spatial databases: Clustering methods (space filling curves), storage methods (R-tree, Grid files).

Assessment breakdown:
75% written examination, 25% project evaluation

University: Université Libre de Bruxelles (ULB)
Department: École polytechnique de Bruxelles
Course ID: DBSA (INFO-H-417)
Course name: Database Systems Architecture
Name and email address of the instructors: Mahmoud Sakr (Mahmoud.Sakr@ulb.be)
Web page of the course: https://www.ulb.be/en/programme/info-h417
Semester: 1
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 24h.
- Exercises: 12h.
- Projects: 24h.

Goals:

In contrast to a typical introductory course in database systems where one learns to design and query relational databases, the goal of this course is to get a fundamental insight into the implementation aspects of database systems. In particular, we take a look under the hood of relational database management systems, with a focus on query and transaction processing. By having an in-depth understanding of the query-optimisation-and-execution pipeline, one becomes more proficient in administering DBMSs, and hand-optimising SQL queries for fast execution.

Learning outcomes:

Upon successful completion of this course, the student:

- Understands the workflow by which a relational database management systems optimises and executes a query
- Is capable of hand-optimising SQL queries for faster execution
- Understands the I/O model of computation, and is capable of selecting and designing data structures and algorithms that are efficient in this model (both in the context of datababase systems, and in other contexts).
- Understands the manner in which relational database management systems provide support for transaction processing, concurrency control, and fault tolerance

Readings and text books:

- Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer Widom. *Database Systems: The Complete Book*, Prentice Hall, second edition, 2008.
- Raghu Ramakrishnan, Johannes Gehrke. *Database Management Systems*. McGraw-Hill, third edition, 2002.

Prerequisites:

- Introductory course on relational databases, including SQL and relational algebra
- Course on algorithms and data structures
- Knowledge of the Java programming language

Table of contents:

- Query Processing
  With respect to query processing, we study the whole workflow of how a typical relational database management system optimises and executes SQL queries. This entails an in-depth study of:
  - translating the SQL query into a “logical query plan";
  - optimising the logical query plan;
  - how each logical operator can be algorithmically implemented on the physical (disk) level, and how secondary-memory index structures can be used to speed up these algorithms; and
  - the translation of the logical query plan into a physical query plan using cost-based plan estimation.
- Transaction Processing
  - Logging
  - Serializability
  - Concurrency control

Assessment breakdown:

70% written examination, 30% project evaluation

University: Université Libre de Bruxelles (ULB)
Department: École polytechnique de Bruxelles
Course ID: DW (INFO-H-419)
Course name: Data Warehousing
Name and email address of the instructors: Esteban Zimányi (ezimanyi@ulb.ac.be)
Web page of the course: http://cs.ulb.ac.be/public/teaching/infoh419
Semester: 1
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 24h.
- Exercises: 24h.
- Projects: 12h.

Goals:
Relational and object-oriented databases are mainly suited for operational settings in which there are many small transactions querying and writing to the database. Consistency of the database (in the presence of potentially conflicting transactions) is of utmost importance. Much different is the situation in analytical processing where historical data is analyzed and aggregated in many different ways. Such queries differ significantly from the typical transactional queries in the relational model:

1. Typically analytical queries touch a larger part of the database and last longer than the transactional queries;
2. Analytical queries involve aggregations (min, max, avg, ...) over large subgroups of the data;
3. When analyzing data it is convenient to see it as multi-dimensional.

For these reasons, data to be analyzed is typically collected into a data warehouse with Online Analytical Processing support. Online here refers to the fact that the answers to the queries should not take too long to be computed. Collecting the data is often referred to as Extract-Transform-Load (ELT). The data in the data warehouse needs to be organized in a way to enable the analytical queries to be executed efficiently. For the relational model star and snowflake schemes are popular designs. Next to OLAP on top of a relational database (ROLAP), also native OLAP solutions based on multidimensional structures (MOLAP) exist. In order to further improve query answering efficiency, some query results can already be materialized in the database, and new indexing techniques have been developed.

The first and largest part of the course covers the traditional data warehousing techniques. The main concepts of multidimensional databases are illustrated using the SQL Server tools. The second part of the course consists of advanced topics such as data warehousing appliances, data stream processing, data mining, and spatial-temporal data warehousing. The coverage of these topics connects the data warehousing course with and serves as an introduction towards other related courses in the program. Several associated partners of the program contribute to the course in the form of invited lectures, case studies, and "proof of technology" sessions.

Learning outcomes:
At the end of the course students are able to

- Explain the difference between operational databases and data warehouses and the necessity of maintaining a dedicated data warehousing system
- Understand the principles of multidimensional modeling
- Design a formal conceptual multidimensional model based on an informal description of the available data and analysis needs
- Implement ETL-scripts for loading data from operational sources a the data warehouse
- Deploy data cubes and extract reports from the data warehouse
- Explain the main technological principles underlying data warehousing technology such as indexing and view materialization.

Readings and text books:

- Matteo Golfarelli, Stefano Rizzi. *Data Warehouse Design: Modern Principles and Methodologies.* McGraw-Hill, 2009
- Christian S. Jensen, Torben Bach Pedersen, Christian Thomsen. *Multidimensional Databases and Data Warehousing.* Morgan and Claypool Publishers, 2010
- Ralph Kimball, Margy Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, third edition, Wiley, 2013.
- Esteban Zimányi, Alejandro Vaisman. *Data Warehouse Systems: Design and Implementation.* Springer,

2014

Prerequisites:

- A first course on database systems covering the relational model, SQL, entity-relationship modelling, constraints such as functional dependencies and referential integrity, primary keys, foreign keys.
- Data structures such as binary search trees, linked lists, multidimensional arrays.

Table of contents:

There is a mandatory project to be executed in three steps in groups of 3 students, using the tools learned during the practical sessions, being SQL Server, SSIS, SSAS, and SSRS. Below is the succinct summary of the theoretical part of the course:

- Foundations of multidimensional modelling
- Dimensional Fact Model
- Querying and reporting a multidimensional database with OLAP
- Methodological aspects for data warehouse development
- Populating a data warehouse: The ETL process
- Using the data warehouse: data mining and reporting

Assessment breakdown:

75% written examination, 25% project evaluation

University: Université Libre de Bruxelles (ULB)
Department: École polytechnique de Bruxelles
Course ID: WM (INFO-H-420)
Course name: Management of Data Science and Business Workflows
Name and email address of the instructors: Dimitris Sacharidis (dimitris.sacharidis@ulb.be)
Web page of the course: https://www.ulb.be/en/programme/info-h420
Semester: 1
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 24h.
- Exercises: 24h.
- Assignments: 12h.

Goals:

This course introduces basic concepts for managing workflows in data science applications and business processes. The first part of the course focuses on *business process management* and considers identification, modeling, analysis, simulation, redesign, and mining based on the Business Process Modeling and Notation (BPMN) workflow language. The second part focuses on *data science workflows* and discusses modeling, execution, and optimization, and also introduces various topics on *responsible data science*.

Learning outcomes:

At the end of the course students are able to:

- Explain the business process management cycle.
- Design a formal model of the business process based on an informal description.
- Identify opportunities for optimizing business processes.
- Describe data science workflows.
- Identify the costs associated with executing data science workflows.
- Optimize data science workflows.
- Identify concerns about data privacy and bias.
- Propose techniques to increase the explainability of data science workflows.

Readings and text books:

- Dumas, La Rosa, Mendling & Reijers. *Fundamentals of Business Process Management* (second edition), Springer 2018.

Prerequisites:

- Basic programming understanding.
- Basic set theory (notions such as set, set operations, sequence, multiset, function) and logics (mathematical notation and argumentation; basic proofs).
- Basic graph theory (notions such as graphs, reachability, transitivity).

Table of contents:

There are four assignments to be realized in groups.

Below is a high-level overview of the theoretical part of the course:

- Business Process Management
  - Short overview of business processes and the need to manage them.
  - Describing business processes, modeling the control flow, data and resource perspectives.
  - Analysis of business processes, qualitatively and quantitatively.
  - Redesign of business processes.
  - Mining Process Logs.
- Data Science Workflows
  - Short overview of data science workflows.
  - Describing workflows in data science.
  - Analysis and optimization of data science workflows.
  - Data privacy.
  - Explainability of data science workflows.
  - Bias and fairness in data science workflows.

Assessment breakdown:
40% written examination, 60% assignment evaluation

University: Université Libre de Bruxelles (ULB)
Department: École polytechnique de Bruxelles
Course ID: DM (INFO-H-423)
Course name: Data Mining
Name and email address of the instructors: Mahmoud Sakr (Mahmoud.Sakr@ulb.be)
Web page of the course: http://cs.ulb.ac.be/public/teaching/infoh423
Semester: 1
Number of ECTS: 5

Course breakdown and hours:

- Lectures: 24h.
- Exercises: 24h.
- Projects: 12h.

Goals:

Data Mining aims at finding useful regularities in large structured and unstructured data sets. The goal of this course is to get a fundamental understanding of popular data mining techniques strengths and limitations, as well as their associated computational complexity issues. It will also identify industry branches which most benefit from Data Mining such as health care and e-commerce. The course will focus on business solutions and results by presenting case studies using real (public domain) data. The students will use recent Data Mining software.

Learning outcomes:

At the end of the course students are able to

- Establish the main characteristics and limitations of algorithms for addressing data mining tasks.
- Select, based on the description of a data mining problem, the most appropriate combination of algorithms to solve it.
- Develop and execute a data mining workflow on a real-life dataset to solve a data-driven analysis problem.
- Use recent data mining software for solving practical problems.
- Identify promising business applications of data mining.

Readings and text books:

- David J. Hand, Heikki Mannila, Padhraic Smyth. *Principles of Data Mining*. MIT Press, 2001.
- Delmater Rhonda, Hancock Monte. *Data Mining Explained*. Digital Press, 2001.
- Pang-Ning Tan, Michael Steinbach, Vipin Kumar. *Introduction to Data Mining*. Pearson Education (Addison Wesley), 2006.

Prerequisites:

- Programming experience
- Data structures
- Algorithms, complexity

Table of contents:

- Data mining
  - Data mining and knowledge discovery
  - Data mining functionalities
  - Data mining primitives, languages, and system architectures
- Data preprocessing
  - Data cleaning
  - Data transformation
  - Data reduction
  - Discretization and generating concept hierarchies
- Data mining algorithms
  - Motivation and terminology
  - Different algorithm types
- Classification and Clustering
  - Classification: SVM classifier
  - Clustering: K-means, Latent Semantic Analysis

- Mining Associations and Correlations
  - Item sets
  - Association rules
  - Generating item sets and rules efficiently
  - Correlation analysis
- Advanced techniques, data mining software, and applications
  - Text mining: extracting attributes (keywords), structural approaches (parsing, soft parsing).
  - Bayesian approach to classifying text
  - Web mining: classifying web pages, extracting knowledge from the web
  - Recommendation systems
  - Data mining software and applications

Assessment breakdown:
75% written examination, 25% project evaluation

Chapter 3

UPC — Advanced Data Management and Analysis Foundations (2nd semester)

3.1 Detailed Description of Curriculum

This semester focuses on fundamental concepts to become a proficient data-driven AI engineer. Accordingly, it aims at training students in understanding how data management can scale to large volumes while potentially also dealing with dealing with high-throughput and heterogeneous (multimodal) data. This semester also covers fundamental concepts to conduct rigorous data analysis with Machine Learning techniques and capabilities to use high-performance infrastructures for complex data analytics workflows and its distribution in either distributed or high-performance systems. Additionally, the students will mandatorily take an ethics course and will learn how to autonomously study a new emerging field on their own. This semester has a strong practical component, which is reflected on realistic projects during the semester.

University: Universitat Politècnica de Catalunya (UPC)
Department: Department of Service and Information System Engineering
Course ID: BDM
Course name: Big Data Management
Name and email address of the instructors: Álex Barceló (alex.barcelo@upc.edu)
Web page of the course: https://www.fib.upc.edu/en/studies/masters/master-innovation-and-research-informatics/curriculum/syllabus/BDM-MIRI
Semester: 2
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 27 h.
- Laboratories: 27 h.
- Self-Study: 96 h.

Goals:
The main goal of this course is to analyze the technological and engineering needs of Big Data Management. The enabling technology for such a challenge is the Cloud, which provide the elasticity needed to properly scale the infrastructure as the needs of the company grow. Thus, students will learn advanced data management techniques (i.e., NOSQL solutions) that also scale with the infrastructure. Being Big Data Management the evolution of Data Warehousing, such knowledge (see the precious semester courses) is assumed in this course, which will specifically focus on the management of large and high-throughput data.

On the one hand, to deal with high volumes of data, we will see how a distributed file system can scale to as many machines as necessary. Then, we will study different physical structures we can use to store our data in it. Such structures can be in the form of a file format at the operating system level, or at a higher level of abstraction. In the latter case, they take the form of either sets of key-value pairs, collections of semi-structured documents or column-wise stored tables. We will see that, independently of the kind of storage we choose, current highly parallelizable processing systems using functional programming principles, whose processing framework can rely on temporal files but mostly in in-memory structures (like Spark).

On the other hand, to deal with high velocity of data, we need some low latency system which processes either streams or micro-batches. However, nowadays, data production is already beyond processing technologies capacity. More data is being generated than we can store or even process on the fly. Thus, we will recognize the need of (a) some techniques to select subsets of data (i.e., filter out or sample), (b) summarize them maximizing the valuable information retained, and (c) simplify our algorithms to reduce their computational complexity (i.e., doing one single pass over the data) and provide an approximate answer.

Finally, the complexity of complex data ecosystems (combining all the necessary tools in a collaborative ecosystem), which typically involves several people with different backgrounds, requires the definition of a high level architecture that abstracts technological difficulties and focuses on functionalities provided and interactions between modules. Therefore, we will also present good practices for DataOps / MLOps and the concept of data governance.

Learning outcomes:
Upon successful completion of this course, the student is able to:

- Knowledge
  - Understand the main advanced methods of data management and design and implement non-relational database managers, with special emphasis on distributed systems.
  - Understand, design, explain and carry out parallel information processing in massively distributed systems.
  - Manage and process a continuous flow of data.
  - Design, implement and maintain complex data ecosystems that manage the complete data life cycle in analytical environments.
- Skills
  - Design distributed NOSQL databases.
  - Produce a functional program to process large datasets in distributed environments.
  - Manage and process a stream of data.
  - Design complex data ecosystems.

Readings and text books:

- M. Tamer Özsu, Patrick Valduriez. *Principles of Distributed Database Systems*, Springer, 2011.
- Ling Liu, M. Tamer Özsu. *Encyclopedia of Database Systems*, Springer, 2009.

- Pramod J. Sadalage, Martin Fowler. *NoSQL Distilled*, Addison-Wesley, 2013.
- Mark Grover, Ted Malaska, Jonathan Seidman, Gwen Shapira. *Hadoop Application Architectures*, O'Rilley, 2015.
- Hasso Plattner, Alexander Zeier. *In-Memory Data Management*, Springer, 2011.
- Matei Zaharia. *An Architecture for Fast and General Data Processing on Large Clusters*, ACM Press, 2016.
- Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman. *Mining of Massive Datasets*, http://www.mmds.org/
- Charu C. Aggarwal (Ed.). *Data Streams*, Springer, 2007.

Prerequisites:

- Advanced Databases (ADB)
- Database Systems Architecture (DBSA)
- Data Warehousing (DW)

Table of contents:

Assessment breakdown:
60% written examination, 40% lab marks

University: Universitat Politècnica de Catalunya
Department: Department of Service and Information System Engineering
Course ID: SDM
Course name: Semantic Data Management
Name and email address of the instructors: Anna Queralt (anna.queralt@upc.edu)
Web page of the course: https://www.fib.upc.edu/en/studies/masters/erasmus-mundus-master-big-data-management-and-analytics/curriculum/syllabus/SDM-BDMA
Semester: 2
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 27 h.
- Laboratories: 27 h.
- Self-study: 96 h.

Goals:
Multimodal and data heterogeneity is a key aspect for Data Science and data-driven Artificial Intelligence, how to tackle data variety in real-world projects is yet not clear and there are no standarized solutions for this challenge.

In this course the student will be introduced to advanced database technologies, modeling techniques and methods for tackling variety. We will also explore the difficulties that arise when combining Variety with large volumes of data and / or high-throughput data. The focus of this course is on the need to enrich the available data (typically owned by the organization) with external data (typically non-structured or semi-structured), in order to gain further insights into the organization business domain. For example, data coming from social networks such as Facebook or Twitter; data released by governmental bodies (such as town councils or governments); data coming from sensor networks (such as those in the city services within the Smart Cities paradigm); etc.

The student will learn about semantic-aware data management as the most promising solution for this problem. As such, it will be introduced to graph modeling, storage and processing. Special emphasis will be paid to semantic graph modeling (i.e., ontology languages such as RDF and OWL) and its specific storage and processing solutions.

Learning outcomes:
The student will learn models, languages, methods and techniques to cope with Variety in the presence of Volume and Velocity. Specifically:

- Graph modeling, storage and processing,
- Semantic Models and Ontologies (RDF, RDFS and OWL),
- Logics-based foundations of ontology languages (Description Logics) and
- Specific storage and processing techniques for semantic graphs.

The students will learn how to apply the above mentioned foundations to automate the data management lifecycle in front of Variety. We will focus on semantic data governance protocols and the role of semantic metadata artifacts to assist the end-user.

Readings and text books:

- Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, Pierre Senellart, *Web Data Management*, Cambridge University Press, 2011.
- Franz Baader, Diego Calvanese, Deborah L. McGuinness, Daniele Nardi, Peter F. Patel-Schneider, *The Description Logic Handbook*, Cambridge University Press, 2010.
- Sven Groppe, *Data Management and Query Processing in Semantic Databases*, Springer, 2011.

Prerequisites:

- Advanced Databases (ADB)

Table of contents:

- Introduction:
  - Variety and Variability. Definition. External (non-controlled) data sources. Semi-structured data models, non-structured data. Schema and data evolution.
  - The data management life-cycle. Challenges when incorporating external (to the organisation) semi-structured and non-structured data.
- Foundations:

  - Graph management: the graph data model, graph databases, graph processing.
  - Semantic-aware management: Ontology languages (RDF, RDFS, OWL). Logical foundations and Description Logics. Storage. Processing (SPARQL).
- Applications:
  - The Semantic Data Lake. Semantic annotations and semantic data governance.
  - Semantic-aware metadata artifacts to automate data integration, linkage and / or cross of data between heterogeneous data sources.

Assessment breakdown:
40% written examination, 60% exercises and laboratories

University: Universitat Politècnica de Catalunya
Department: Department of Computer Science
Course ID: ML
Course name: Machine Learning
Name and email address of the instructors: Marta Arias (marias@cs.upc.edu)
Web page of the course: https://www.fib.upc.edu/en/estudis/masters/master-en-ciencia-de-dades/pla-destudis/assignatures/ML-MDS
Semester: 2
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 27 h.
- Laboratories: 27 h.
- Self-study: 96 h.

Goals:

The aim of machine learning is the development of theories, techniques and algorithms to allow a computer system to modify its behavior in a given environment through inductive inference. The goal is to infer practical solutions to difficult problems –for which a direct approach is not feasible– based on observed data about a phenomenon or process. Machine learning is a meeting point of different disciplines: statistics, optimization and algorithmics, among others.

The course is divided into conceptual parts, corresponding to several kinds of fundamental tasks: supervised learning (classification and regression) and unsupervised learning (clustering, density estimation). Specific modelling techniques studied include artificial neural networks and support vector machines. An additional goal is getting acquainted with python and its powerful machine learning libraries.

Learning outcomes:

- Formulate the problem of (machine) learning from data, and know the different machine learning tasks, goals and tools.
- Organize the workflow for solving a machine learning problem, analyzing the possible options and choosing the most appropriate to the problem at hand.
- Ability to decide, defend and criticize a solution to a machine learning problem, arguing the strengths and weaknesses of the approach. Additionally, ability to compare, judge and interpret a set of results after making a hypothesis about a machine learning problem.
- To be able to solve concrete machine learning problems with available open-source software.

Readings and text books:

- C.M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.
- V.S. Cherkassky, F. Mulier, *Learning from data: concepts, theory, and methods*, John Wiley, 2007.
- E. Alpaydin, *Introduction to machine learning*, The MIT Press, 2014.
- K.P. Murphy, *Machine learning: a probabilistic perspective*, MIT Press, 2012.

Prerequisites:

- Elementary notions of probability and statistics.
- Elementary linear algebra and real analysis.
- Good programming skills in a high-level language.

Table of contents:

- Introduction to Machine Learning.
  General information and basic concepts. Overview to the problems tackled by machine learning techniques. Supervised learning (classification and regression), unsupervised learning (clustering and density estimation) and semi-supervised learning (reinforcement and transductive). Examples.
- Supervised machine learning theory.
  The supervised Machine Learning problem setup. Classification and regression problems. Bias-variance tradeoff. Regularization. Overfitting and underfitting. Model selection and resampling methods.
- Linear methods for regression.
  Error functions for regression. Least squares: analytical and iterative methods. Regularized least squares. The Delta rule. Examples.
- Linear methods for classification.
  Error functions for classification. The perceptron algorithm. Novikoff's theorem. Separations with maxi-

mum margin. Generative learning algorithms and Gaussian discriminant analysis. Naive Bayes. Logistic regression. Multinomial regression.

- Artificial neural networks.
  Artificial neural networks: multilayer perceptron and a peak into deep learning. Application to classification and to regression problems.
- Kernel functions and support vector machines.
  Definition and properties of Kernel functions. Support vector machines for classification and regression problems.
- Unsupervised machine learning.
  Unsupervised machine learning techniques. Clustering algorithms: EM algorithm and k-means algorithm.
- Ensemble methods.
  Bagging and boosting methods, with an emphasis on Random Forests.

40% final exam, 40% practical work and 20% mid-term exam.

University: Universitat Politècnica de Catalunya (UPC)
Department: Department of Computer Architecture
Course ID: HPDA
Course name: High-Performance Data Analytics
Name and email address of the instructors: Josep-Lluís Berral (josep.ll.berral@upc.edu)
Web page of the course: New course to be created for DEAI
Semester: 3
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 25 h.
- Interactive exercises: 10 h.
- Laboratory exercises: 25 h.
- Self-study, assignments and research: 90 h.

Goals:
Underlying architectures and computational resources have great impact on Artificial Intelligence, Machine Learning and Data Science processes. Predictive analytics for Big Data require high-performance infrastructures, such as computing clusters, supercomputers and Cloud deployments. And in its turn, such infrastructures can enable processes through acceleration, distribution and cooperation. Building algorithms aware of the architecture allows us to scale up/out analytics, modelling and prediction; accelerate training methods through GPU/TPU/DPU specialized devices for data operations; and federate data processing with special concern on privacy and security. Furthermore, it allows us to adapt our algorithms to work from high-performance resources (i.e., in the Cloud, next to high-performance) to low-power devices (i.e., in the Edge, next to the user). This subject focuses on discovering and using such resources towards efficient data analytics and AI.

This course will focus on students learning and being able to create solutions leveraging four state-of-art concepts: 1) high-performance computing technologies and architectures for data analytics and AI, such as supercomputers; 2) acceleration technologies for AI, such as GPU-based environments; 3) distributed analytics and AI across heterogeneous architectures, such as high-performance vs low-power devices in Cloud-Edge architectures; and 4) federated learning architectures for privacy-preserving ML processes, from the architecture perspective.

Learning outcomes:
Upon successful completion of this course, the student will be able to:

- Prepare data analytics and AI processes for supercomputing resources and infrastructures, and adapt pipelines for their execution in a high-performance computing environment.
- Leverage acceleration technologies for massive data processing such as GPUs, through specific programming models oriented towards AI fundamental operations.
- Design and deploy distributed architectures for data analytics and AI, across heterogeneous infrastructures with different computational capabilities and requirements.
- Prepare federated learning architecture and algorithms, considering the distribution of data, models and participants.

Readings and text books:

- John L. Hennessy, David A. Patterson. "Computer architecture : a quantitative approach". Elsevier/Morgan Kaufmann. [2019]. ISBN : 9780128119051.
- Thomas Sterling, Matthew Anderson, Maciej Brodowicz. "High performance computing : modern systems and practices". Morgan Kaufmann, [2018]. ISBN: 9780124201583.
- Maarten van Steen, Andrew S. Tanenbaum. "Distributed Systems", 4th ed.. Distributed-systems.net. [2023]. ISBN: 978-9081540636.
- Heiko Ludwig, Nathalie Baracaldo. "Federated Learning. A Comprehensive Overview of Methods and Applications". Springer Cham. [2022]. ISBN 9783030968953.

Other Literature:

- Federated Learning: Privacy and Incentive. URL: https://federated-learning.org/FLPI/
- Yuan Tang. "Distributed Machine Learning Patterns". Manning. [2023]. ISBN: 9781617299025.
- BSC MareNostrum-V Documentation. URL: https://www.bsc.es/supportkc/docs/MareNostrum5/intro/

Prerequisites:

- Elementary notions of machine learning and data analytics processes and algorithms.
- Technical knowledge on UNIX systems and networks, also distributed systems.
- Basic programming skills in Python and C language.

Table of contents:

- Introduction to HPC for DS & AI
  - Cloud, clusters and supercomputers. Introduction to high-performance infrastructures for scientific computing and industry.
  - Resource management and energy consumption. Computational resources, operational costs and energy consumption for DS and AI applications.
  - Performance and Quality of Service. Concept understanding of performance, throughput and quality of service in finite-resource systems.
- Accelerators and Enablers
  - SIMD operations. Introduction to "Single Instruction, Multiple Data" operations, intrinsic vectorization and matrix operations.
  - Graphic-Processing Units. Architecture of GPUs towards AI, and introduction to CUDA programming for AI operations.
- Distributed Machine Learning
  - Parallelism for data analytics and AI. Multi-processing, multi-machine and distributed computing for AI algorithms.
  - Parallel programming models for HPC. Introduction to OpenMP and MPI towards AI algorithms and scalable systems.
  - Partition of ML algorithms. Considerations, methods and technologies, aware of data and workload locality.
- Applied Federated Learning
  - Federated and Swarm Learning. Architectures, algorithms and scenarios. Protection against free-riders, noise and poisoning.
  - Applied cases and examples. Medical uses for federated learning and distributed analytics.

Assessment breakdown:

The assessment is based on the acquired knowledge on the fundamentals on theory, methodology and technology (*Theory*); the capacity to develop and deploy such knowledge in hands-on and practical exercises (*Exercises*); and the participation in class with regards to discussion, research about topics commented in class, and presentations (*Participation*).

The final mark will be 40% final exam on *Theory*, 40% on the assignments as *Exercises*, and 20% on active *Participation*.

University: Universitat Politècnica de Catalunya (UPC)
Department: Department of Service and Information System Engineering
Course ID: DEAIS
Course name: Data Engineering and Artificial Intelligence Seminars
Name and email address of the instructors: Oscar Romero (oscar.romero@upc.edu)
Web page of the course: To be created
Semester: 2
Number of ECTS: 1

Course breakdown and hours:

- Lectures: 21 h.
- Autonomous work: 129 h.

Goals:
In these seminars, students get a view of recent developments in Data Engineering and Artificial Intelligence. Invited lecturers from associated partners, both from academia and industry, will cover business cases, research topics, internships and master's thesis subjects, to present novel aspects in the field and also motivate the three specialisations according to their interests. Students conduct a state-of-the art research in one of the topics, which is jointly evaluated by all partners in the summer school.

Learning outcomes:
Upon successful completion of this course, the student is able to:

- Read and understand scientific papers
- Develop critical thinking when assessing scientific papers
- Write and explain a state-of-the-art in a rigorous manner
- Elaborate on recent trends in Data Engineering and Artificial Intelligence

Readings and text books:

- Gordana Dodig-Crnkovic, *Theory of Science*, online resource: http://www.idt.mdh.se/kurser/ct3340/archives/ht04/theory_of_science_compendiu.pdf
- Jennifer Widom, *Tips for Writing Technical Papers*, online resource: https://cs.stanford.edu/people/widom/paper-writing.html
- Robert Siegel, *Reading Scientific Papers*, online resource: https://web.stanford.edu/~siegelr/readingsci.htm

Prerequisites:

- No prerequisites

Table of contents:
The seminars content will vary from course to course as they will focus on current hot topics in Data Engineering and Artificial Intelligence.

Assessment breakdown:
100% 50% Poster presentation at the summer school.

Note: Besides the poster presentation, attendance to the summer school is mandatory. To be evaluated (see the formula above) students must guarantee >75% overall attendance to the summer school events and the semester seminars.

University: Universitat Politècnica de Catalunya (UPC)
Department: Department of Service and Information System Engineering
Course ID: DDE
Course name: Debates on Data Ethics
Name and email address of the instructors: Petar Jovanovic (petar.jovanovic@upc.edu)
Web page of the course: to be created
Semester: 2
Number of ECTS: 3

Course breakdown and hours:

- Lectures: 18 h.
- Autonomous work: 32 h.

Goals:
This course fosters the social competences of students, by building on their acquired oral communication skills to debate about the ethical consequences related to data-driven initiatives. In this course we cover the Human Agency and Oversight, Societal and Environmental Well-Being as well as Non-Discrimination and Diversity issues that may arise due to unethical use of data. The aim of this course is to develop their critical attitude and reflection. A written summary of their position is meant to train their writing skills.

During the course sessions the debates discussing innovative and visionary ideas on Data Ethics will take place. You must read the available material before the debate. Then, during the debate you will be assign to a group: either to defend an idea, or go against it. You may also be asked to moderate the debate. Then, the debate takes place and afterwards, each group needs to write down a report with their conclusions.

Learning outcomes:
Upon successful completion of this course, the student will develop:

- Ability to study and analyze problems in a critical mood
- Ability to critically read texts
- Develop critical reasoning with special focus on ethics and social impact
- Develop soft skills to defend - criticize a predetermined position in public
- Improve the writing skills

Readings and text books:

- Rudi Volti, *Society and technological change*, Worth, 2009.
- Richard T. De George, *The Ethics of information technology and business*, Blackwell, 2003.
- David Elliott, *Energy, society, and environment: technology for a sustainable future*, Routledge, 2003.
- Kord Davis, *Ethics of big data*, O'Reilly, 2013.

Prerequisites:

- No prerequisites

Table of contents:
Debates on data ethics and its social impact (some examples):

- Ethics codes
- Right to privacy
- Content piracy
- Social responsibility of IT companies
- Artificial Intelligence and their limits

Assessment breakdown:
80% Debate evaluations, 20% Ethical analysis of a data-based business idea

Chapter 4

UniPD — Business, Economics, Financial Data Science (3rd semester)

4.1 Motivation and Background

The School of Science of University of Padua (UniPd) offers a Master Degree in Data Science[1]. The goal is to build data scientists whose solid technical background is complemented by a multidisciplinary preparation on various fields in which Big Data emerges.

On the basis of this vision, the UniPd Degree in Data Science focuses on advanced together statistical and AI methods, such as Deep Learning, and on their application to domains in which a large amount of data need to be analyzed with the goal to extract insightful and actionable patterns. In particular, the degree emphasizes the application for the analytics of human data, economic and financial data, biological data, and on the legal repercussion of these analyses (cf. the recent EU's release of the General Data Protection Regulation - GDPR).

This vision is shared by the Erasmus Mundus in Big Data Management and Analytics (BDMA), which motivates our interest to join the program. In the light of the focus described above, UniPd is interested to join BDMA program by providing a specialization on advanced Data-Science Methods and their application to different contexts. UniPd believes that this will enlarge the expertise of the consortium and provide students with one additional specialization that differs in topics from those already offered. Further details follow in the next section.

The University of Padua (UniPd) has long-standing experience in Erasmus Mundus Joint Master Degrees, having participated in the Erasmus Mundus program since it was launched in 2004.

As per 2020, the University coordinates one Erasmus Mundus: Sustainable TErritorial DEvelopment - STEDE[2]. STEDE is now at its 9th batch and, starting from September 2020, it will be supported independently by the Partner universities thanks to the excellent reputation achieved. It is widely recognized as an example of good practice and has been included in EACEA publications and conferences on EMJMD. In April 2020, an excellent feedback was received by EACEA on the interim report submitted in February 2020.

UniPD is also partner of other six programs, all of which are at their 2nd or 3rd edition, proving solid cooperation mechanisms and quality results.

- MEDFOR - Erasmus Mundus Joint Master Course in Mediterranean Forestry and Natural Resources Management
- NUPHYS - Erasmus Mundus Joint Master Degree in Nuclear Physics
- PLANTHEALTH - European Master degree in Plant Health in Sustainable Cropping Systems
- SUFONAMA - Master in Sustainable Forest and Nature Management
- SUTROFOR - Sustainable Tropical Forestry Erasmus Mundus Masters Course
- TPTI - Techniques, Patrimoines, Territoires de l'Industrie: Histoire, Valorisation, Didactique

Since the start of Erasmus Mundus in 2004, UniPD has also participated in the programmes listed below, some of which are still offered without Erasmus Mundus funding (they are marked with *).

- ALGANT and ALGANT II *
- ASTROMUNDUS
- MBIO - Erasmus Mundus Master of Bioethics
- SAHC - Advanced Master in Structural Analysis of Monuments and Historical Constructions*
- EXTATIC
- FONASO

[1] https://www.unipd.it/en/data-science
[2] https://em-stede.eu/

- FUSION-DOC

4.2 Proposal of Specialization

The proposal of specialization aims to provide students with advanced Data-Science methods and strengthen their background in Statistics and Deep Learning. In this respect, UniPd will provide a first mandatory course on Statistical Inference, and a second mandatory course on Deep Learning, the latter with an emphasis on the analysis of human data (see the course description below). Furthermore, students with a strong interest in statistical methods can opt to further enlarge their background of Data-Science methods by also taking a course on Stochastic Model.

While fortifying their background on Data-Science methods, the specialization puts a strong accent on the applications to Data Science. In particular, students will be able to learn the diverse background knowledge that enables them to carry on data-science projects in different domains, and to extract and understand the results that are being obtained.

In particular, students (1) will learn how to conduct an analysis of data that originate from different contexts, ranging from human, business-process, financial and economic data. Students (2) will acquire the knowledge of these domains to be able to interpret the results. Last but not least, the list of optional courses included a course that aims to introduce non-law students to a proper understanding of the main legal issues related to the processing of data, including those personal.

The specialization will provide a repertoire of optional courses that allow each student to decide the domains where she/he wants to further specialize her/his interest, and/or to strengthen their background in stochastic methods, and/or to deepen their knowledge on the legal and ethical aspects of data-science.

At the end of their specialization curriculum and the final Master project, the student will obtain the degree in Data Science from University of Padua.

The emphasis on Deep Learning, Stochastic and Statistical Methods, and on the application to several data-intensive domains constitute a remarkable, complementary specialization with respect to those already available in the BDMA Erasmus Program. This can improve the desirability of the program, and attract a larger number of high-quality student candidates who are interested in the topics covered by this additional specification.

4.2.1 Specialization Curriculum

Necessary Background Knowledge The specialization assumes students to have previously acquired certain knowledge. In particular, students are expected to possess a basic knowledge of (1) some programming language, preferably Python, (2) Probability Theory, (3) Lineal Algebra, and (4) Machine Learning. The knowledge of Machine Learning acquired during the first year, e.g. through the "Data Mining" course at ULB, is sufficient.

The optional course on Stochastic Methods requires students to possess the basic notions of Differential and Integral calculus. Some of the extensive knowledge on database technology acquired during the semester at ULB is pre-requisite for the optional courses on Bio-informatics.

4.2.2 Academic Calendar and Marking System

The following is the calendar relevant for BDMA, including examination and resit periods:

- Beginning of third semester: ca. 1 October.
- End of the third semester: ca. 15 January.
- Examination period: 15 January – 28 February.
- Master Project: from March to August.
- Examination resit period (upon BDMA consortium approval): June–July.

During the examination period of January and February, each course will be given two examination attempts, typically one in January and the second in February. During the resit periods, one attempt per course will be offered, but students will need to file a request that requires an approval from the BDMA consortium.

Exams are marked with a grade between 0 and 30, with the sufficient passing mark being 18. Outstanding students can be given a mark of "30 e lode", i.e. a cum-laude mark.

4.2.3 University Fees

The annual enrollment fees of University of Padua is € 2600, and they cover the entire academic year, including both the first and the second semester.

4.2.4 Master Thesis

Students will then carry on the final Master project at UniPd under the supervision of one of the professors and lecturers involved in the Data Science master program. The final project will be carried out in collaboration with a company, and will aim to address the company's questions. The students will carry on the projects by applying the knowledge acquired during the first and second year of the curriculum. The project will help student develop the "soft skills" to identify the questions and the opportunities, draw a realistic plan and carry it out. The knowledge acquired during the first year via the UPC's course on "Viability of Business Projects" will be of high importance, and will be applied in a real company's setting. The courses on Big and Semantic Data Management at UPC and those in Data warehousing, Advanced Databases at ULB will provide the crucial, technical foundations to extract, manage and elaborate the data that will be analyzed during the Master project.

4.2.5 Language Courses

Students are offered a semester-long Italian language course during the third semester at the University Language Centre of UniPd. Those students able to arrive at the beginning of September can alternatively take an intensive Italian language course before the third semester starts. In the latter case, students can attend an additional course during the first semester at the special rate of ca. € 150.

The attendance of language courses is not mandatory, but it certainly provides benefit at the time of carrying on the Master project. The page at http://cla.unipd.it/en/courses/italian-courses/ describes the UniPd's offers of Italian language courses.

4.2.6 Student Support and Accommodation

UniPd offers a variety of services to its Erasmus Mundus students, which include:

1. Support by the International Office. Full-time multilingual staff at the international office is fully dedicated to assisting international students before their arrival and during their stay.
2. Accommodation. A place in one of the university Halls of Residence for third and fourth semesters is guaranteed for Erasmus Mundus students.
3. Welcome Days and Events are organized by UniPd in September and February for all international students. So will the UniPd team of the BDMA arrange at least a dedicated welcome event in September and additional event at the end of February. The BDMA students who take UniPd as specialization will also join the annual events that are organized for the other students of the Data-Science Program of Padua
4. Meal Support. Special rates at the University canteens are offered to international students for meals (see also https://www.esupd.gov.it/en/our-main-activities/food-services-canteens).
5. Resident-Permit Support. Dedicated support with regards to visa and residence permit applications is provided by the SAOS desk (see https://www.unipd.it/en/node/1456)

The status of students of UniPd will allow the BDMA students to access to the other services provided to every student of UniPd:

- Disability Service: https://www.unipd.it/en/resources-and-supports-students-disability-or-learning-disabilities.
- Psychological Assistance Service and Psychiatric Consultancy Service provided by the Department of Psychology of Mental Development and Socialization (https://www.dpss.unipd.it/en/).
- University Sports Centre (CUS): http://www.cuspadova.it/.

4.3 Key People

This section summarizes the key people who will be involved in the BDMA Erasmus Mundus program from University of Padua, should UniPd become part of the consortium.

Prof. Massimiliano de Leoni will be the main reference person for UniPd for what concerns BDMA. Namely, he will the representative of UniPd to join the steering committee, and to participate in the selection committee, and he will certainly attend the summer schools and the defences. Massimiliano is an Associate Professor for Computer Science at the University of Padua, Italy. He obtained a Ph.D. in Computer Science at SAPIENZA – University of Rome on a thesis on Business Process Flexibility and Adaptation. Prior the current position, he was an Assistant Professor for Computer Science at Eindhoven University of Technology. His teaching experiences are in the area of Database, Distributed Systems, Business Process Intelligence, Process Mining, and Business Information Systems. He was demo chair at the Business-Process-Management conference in 2018, and general chair of the International Conference in Process Mining for 2020 (ICPM 2020). His research is in the Data-Science area with special emphasis on Artificial Intelligence and Machine Learning, Analytics of Business and Human Data, and Business Intelligence. To this date, his h index is 40.

Prof. Marco Ferrante is the director of the Data-Science Master program at UniPd. Marco will act as liaison between the BDMA program and the Data-Science Master at UniPd. Marco Ferrante received a PhD in Functional Analysis and Applications from the SISSA of Trieste, Italy, in 1993. Currently, he is full professor of Probability Theory at the Department of Mathematics "Tullio Levi-Civita" of UniPd since 2002, and he served as Head of the Department from 2014 until 2018. His research interests include Probability theory, Stochastic processes and their applications to biological models, Stochastic calculus and Information retrieval. He has published more than 35 papers in international Journals on Stochastic differential equation, Malliavin calculus, Applied stochastic processes and Information retrieval.

Mrs. Alessandra Gallarano is the Head of the Projects & Partnership Unit of UniPd International Office, where she coordinates EU-funded projects (including the Erasmus Mundus Joint Master Degrees in which UniPd is partner), international partnerships, and joint programmes. Before joining UniPd in 2016, Alessandra served as a project manager for Sapienza University of Rome, as well as for the Directorate General of Higher Education, Portugal and the Council of Europe, France. Alessandra obtained an Erasmus Mundus Joint Master degree in European Studies from the University of Saint Andrews (United Kingdom), Nova University of Lisbon (Portugal) and University of Perpignan (France). Thanks to her academic and professional experience, Alessandra has a solid knowledge and experience in developing and managing Erasmus Mundus Joint Master Degrees and joint programmes in general. Since 2016, she is also part of the EAIE Joint Programmes Network, acting as trainer for international workshops and sessions.

Mrs. Elisa Zambon is the International Relations officer at the University of Padova since 2009, starting as junior project manager of Erasmus Mundus Action 2 projects, responsible for the implementation of the projects, financial management, incoming and outgoing students, reporting activity and drafting of new proposals. Elisa is now in charge of all international incoming students in the framework of Erasmus Mundus Joint Degrees, Erasmus+ KA107 and Bilateral Agreement.

4.4 Detailed Description of Curriculum

The proposal of specialization aims to provide students with advanced Data-Science methods and strengthen their background in Statistics and Deep Learning. In this respect, University of Padua will provide a first mandatory course on Statistical Inference, and a second mandatory course on Deep Learning. The latter course also focuses on the application of Deep Learning techniques on the analysis of human data. Furthermore, students with a strong interest in statistical methods can opt to further enlarge their background of Data-Science methods by also taking a course on Stochastic Model.

While fortifying their background on Data-Science methods, the specialization puts a strong accent on the applications to Data Science. In particular, students will be able to learn the diverse background knowledge that enables them to carry on data-science projects in different domains, and to extract and understand the results that are being obtained.

In particular, students (1) will learn how to conduct an analysis of data that originate from different contexts, ranging from human and economic data, till biological data, and (2) will acquire the knowledge of these domains to be able to interpret the results.

University: Universitá degli Studi di Padova (UniPd)
Department: Dipartimento di Matematica
Course ID: StatLearn
Course name: Statistical Learning
Name and email address of the instructors: Bruno Scarpa (bruno.scarpa@unipd.it)
Web page of the course: https://en.didattica.unipd.it/off/2025/LM/SC/SC2738/003PD/SCP7079227/N0
Semester: 1
Number of ECTS: 6

Lectures, Exercises: 48h.

Goals:
Become familiar with statistical thinking; gain adequate proficiency in the development and use of standard statistical inference tools; be able to analyse datasets using a modern programming language such as R

Learning outcomes:
At the end of the course students should show knowledge of the key concepts, skills in the analysis of data and competency in applications.

Readings and text books:

- Azzalini, A., Scarpa, B., Data analysis and data mining: an introduction. –: Oxford University Press, 2012. https://azzalini.stat.unipd.it/Book-DM/
- Lavine, M., Introduction to Statistical Thought. –: None, 2013. http://people.math.umass.edu/~lavine/Book/book.html
- James, G., Witten, D. Hastie, T., Tibshirani, R., Introduction to Statistical Learning: with Applications in R. New York, NY: Springer US, 2021. https://www-bcf.usc.edu/~gareth/ISL/
- Hastie, T., Tibshirani, R., and Friedman, J., The Elements of Statistical Learning: Data Mining, Inference, and Prediction. –: Springer, 2001. https://web.stanford.edu/~hastie/ElemStatLearn/

Additional notes on the applications can be found in:

- Nolan, D.A. & Speed, T. (2000). Stat Labs: Mathematical Statistics Through Applications. Springer.
- Torgo, L. (2011). Data Mining with R: Learning with Case Studies. Chapman & Hall/CRC.

Methods for specific fields of applications can be found in the following books:

- Campbell, R.C. (1989). Statistics for Biologists (3rd ed.). Cambridge University Press.
- Devore, J.L. (2000). Probability and Statistics for Engineering and the Sciences (5th ed.). Duxbury Press, Pacific Grove, CA.
- Agresti, A. & Finlay. B. (2007). Statistical Methods for the Social Sciences (4th ed.). Prentice Hall

Prerequisites:

- Basic probability theory; multivariable calculus; linear algebra; basic computing skills.

Table of contents:

- Data: summary statistics, displaying distributions; exploring relationships
- Estimation: point estimation; the sampling distribution of an estimator; accuracy of estimation; interval estimation
- Hypothesis testing
- Likelihood: the likelihood, likelihood for several parameters
- Estimation: maximum likelihood estimation; properties of maximum likelihood estimates

Assessment breakdown:
Written examination, and project work

University: Universitá degli Studi di Padova (UniPd)
Department: Dipartimento di Matematica
Course ID: DeepLearn
Course name: Deep Learning and Human Data Analytics
Name and email address of the instructors: Michele Rossi (michele.rossi@unipd.it)
Web page of the course: https://en.didattica.unipd.it/off/2025/LM/SC/SC2738/003PD/SCQ4106915/N0 (This course is shown under a different name, but is renamed as intended in this syllabus for BDMA)
Semester: 1
Number of ECTS: 6

Course breakdown and hours: Lectures, Exercises and Project: 48h.

Goals:
Introduce advanced machine- and deep-learning techniques, and apply them to real-world problem within the human-data domain

Learning outcomes: At the end of the course students will acquire skills related to:

1. the main clustering algorithms for vector data, their pros and cons, their evaluation metrics
2. the main unsupervised learning techniques for unsupervised vector quantization, their performance, advantages and their use within real problems involving biosignals
3. the main modeling techniques for multivariate time series and their use within selected applications in the human data domain
4. the principles and main algorithms for supervised learning through neural networks (feed forward, convolutional and recurrent), their programming in Python, and their use to solve real world problems
5. the main application domains for the techniques at the previous points 1, 2, 3 and 4 and how they can be exploited to tackle relevant problems within the "human data" domain
6. comprehending, selecting and knowing how to use the techniques at the previous points 1, 2, 3 and 4
7. solving a real world problem involving human data analysis, through summarizing its solution through a professional written report, presenting the work done via a conference-style talk, also showcasing the software written for the project
8. implementing and use the techniques at points 1, 2, 3 and 4 in Python

Readings and text books:

- Bishop, Christopher M., Pattern recognition and machine learning. Christopher M. Bishop. New York: Springer, 2006.
- Bengio, Yoshua; Courville, Aaron; Goodfellow, Ian, Deep Learning. Cambridge: MIT Press, 2016.
- Watt, Jeremy; Borhani, Reza; K. Katsaggelos, Aggelos, Machine Learning Refined: Foundations, Algorithms, and Applications. New York: Cambridge University Press, 2020
- Technical reports, scientific papers

Prerequisites: Prior knowledge on Calculus and Linear Algebra, and some basic computer programming is useful, preferably in Python.

Table of contents:

- Vector quantization (VQ):K-means, soft K-means, Expectation Maximization (EM), X-means, DBSCAN (density based clustering)
- Unsupervised VQ algorithms: Self-Organizing Maps (SOM), Gas Neural Networks (GNG): theory and algorithms
- Application to quasi-periodic biometric signals (ECG): Signal pre-processing, normalization, segmentation, efficient representation of ECG, final system design and numerical results
- Deep Neural Networks: gradient descent, Feed Forward Neural Networks: models, training, backpropagation, Convolutional Neural Networks (CNN), Autoencoder architectures for unsupervised feature extraction, denoising autoencoders Recurrent neural networks (RNN) for the analysis of temporal sequences, Residual convolutional neural networks (ResNets), Attention mechanisms for CNN (images)
- Use of Keras and TensorFlow frameworks in Python for the implementation and training of neural networks.
- Training of neural network structured on real datasets

Assessment breakdown:
The student examinations will be carried out through a project. The grade is determined on the basis of a written report and the oral discussion of the project.

University: Universitá degli Studi di Padova (UniPd)
Department: Dipartimento di Matematica
Course ID: TimeSeries
Course name: Time-Series Analysis for Business Economic and Financial Data
Name and email address of the instructors: Mariangela Guidolin (mariangela.guidolin@unipd.it)
Web page of the course: https://en.didattica.unipd.it/off/2022/LM/SC/SC2377/002PD/SCQ0093723/N0
Semester: 1
Number of ECTS: 6

Course breakdown and hours:

- Lectures, and Exercises: 48h.

Goals:
The course aims at introducing the students to the main statistical features and concepts underlying the analysis of data collected over time, as well as providing the basic statistical solutions to analyse such data in economic, financial and business settings. The course comprises technical-practical lessons with R.

Learning outcomes:
At the end of the course, students will learn basic characteristics related to business, economic and financial data, alsothrough illustrative examples.

Students will also acquire skills related to discovering linear and non linear regression models, such as regression splines, local regression, generalized additive models (GAM) for non-linear regression models, as well as tree-based methods (Regression trees, Bagging, Boosting). For the analyis of time-series data, students will also acquire skills to learn how to perform exponential smoothing and ARIMA models.

Readings and text books:

- Notes prepared by instructor will be made available throughout the course.
- Bishop, Christopher M., Pattern recognition and machine learning. Christopher M. Bishop. New York: Springer, 2006.
- James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert, An introduction to statistical learning: with applications in R. New York: Springer, 2013.
- Hyndman, R. J., Athanasopoulos, G., Forecasting: principles and practice.. –: OTexts., 2018.
- Azzalini, A., Scarpa, B., Data analysis and data mining: An introduction.. USA: OUP, 2012.

Prerequisites: Some basic computer programming is useful.

Table of contents:

- General introduction to business, economic and financial data
- Preliminary concepts and illustrative examples
- Moving beyond the linear model
- Linear regression model: main ideas and assumptions
- Nonlinear regression models for new product growth
- Beyond linearity: regression splines, local regression, generalized additive models (GAM)
- Tree-based methods: Regression trees, Bagging, Boosting
- Time series analysis: Exponential Smoothing and ARIMA models

Assessment breakdown:
The student examinations will be carried out through a written and an oral exam.

University: Universitá degli Studi di Padova (UniPd)
Department: Dipartimento di Matematica
Course ID: DataBPI
Course name: Business Process Intelligence
Name and email address of the instructors: Massimiliano de Leoni (massimiliano.deleoni@unipd.it)
Web page of the course: https://en.didattica.unipd.it/off/2025/LM/SC/SC2738/003PD/SCQ5110305/N0
Semester: 1
Number of ECTS: 6

Course breakdown and hours:

- Lectures, Exercises, and Laboratory Sessions: 48h.

Goals:
The course aims to introduce the basic concepts of modelling, analysis and improvement of business processes, on the basis of transaction data related to historical process executions.

Learning outcomes:
Students will learn to model business processes using both Petri net and BPMN notations, analyze their structure, and verify their correctness. They will develop a solid understanding of process mining concepts, apply basic and advanced discovery techniques, perform conformance checking through token replay and alignments, and conduct social network analysis. Moreover, they will learn to incorporate decision, time, and resource perspectives into their analyses. The course will also cover the principles of business process simulation, including the definition and execution of simulation scenarios while ensuring statistical significance. Throughout the course, students will apply these principles in practice using both open-source and commercial tools for process modeling, analysis, and simulation on real-life process data.

Readings and text books:

- An exercise bundle
- Lecture notes on the laboratory sessions
- van der Aalst, Wil M. P. : van der, Modeling business processes a Petri net-oriented approach. Cambridge: MIT Press, 2011.
- van der Aalst, Wil M. P.: Process Mining Data Science in Action. Berlin, Heidelberg: Springer Berlin, 2016.
- Marlon Dumas, Marcello La Rosa, Jan Mendling, Hajo A. Reijers, Fundamentals of business process management. Berlin: Springer, 2018.

Prerequisites: Basic knowledge of logics, algorithms, data structure and programming.

Table of contents:

- MODELING VIA PETRI NETS: Basic concepts of Petri nets, Usage of Petri Nets to model business processes, Structural analysis of Petri Nets, Soundness of business process models, Usage of Woped as software tool to model and check the soundness of process models.
- PROCESS MINING: Introduction to Process Mining and Event Logs, Basic Techniques for Process Discovery and Limitations, Advanced Techniques for Process Discovery (Heuristic Miner and Region Miner), Conformance checking based on token replay, Conformance checking based on alignments, Mining the Additional Perspectives on Decision, Time and Resource, Social Network Analysis, Usage of open-source and commercial software tools for the analysis of business processes through Process Mining.
- BUSINESS PROCESS SIMULATION: Principle and Methodologies for simulation: definitions of warm-up and cool-down intervals, management of the simulation stochasticity via multiple, independent runs, analysis of the outcome, Definitions of the elements of a scenario of business process simulation (arrival rate, branching probabilities, resource pools, activity-duration probability distributions), Usage of BPMN for the specification of business process simulation models, Use of Process Mining to mine business process simulation models, Methodologies for the definition of "What-if" scenarios, Usage of BIMP as software tool to simulate business processes

Assessment breakdown:
Written Exam and Project

University: Universitá degli Studi di Padova (UniPd)
Department: Dipartimento di Matematica
Course ID: LawData
Course name: Law and Data
Name and email address of the instructors: Fiorella Dal Monte (fiorella.dalmonte@unipd.it)
Web page of the course: https://en.didattica.unipd.it/off/2025/LM/SC/SC2738/003PD/SCP7079399/N0
Semester: 1
Number of ECTS: 6

Course breakdown and hours:

- Lectures, Seminars and Workshops: 48h.

Goals:
The course aims to introduce non-law students to a proper understanding of the main legal issues related to the processing of data, personal and non. The first part of the course aims to enable students to approach EU personal data protection regulation. In the second part, instead, students will reflect on the main problems related to the use of data-intensive technologies (big data and artificial intelligence) and the technical and legal solutions now debated.

Learning outcomes:
The students will learn the EU regulations on personal data protection and the legal implications, as well as they will be taken to reflect on the issues related to using AI and Data-Science techniques and on the consequent solutions currently being debated.

Readings and text books:

- Mireille Hildebrandt (2020). Law for Computer Scientists and Other Folk, OUP (available: open access) – especially chapters 2-3-4-5-9-10
- Paul Voigt, Axel von dem Bussche (2017). The EU General Data Protection Regulation (GDPR). A practical guide, Springer
- European Fundamental Rights Agency (2018). Handbook on European data protection law, Luxemburg
- Karen Yeung, Martin Lodge (2019). Algorithmic Regulation, OUP (Public Law Dept. Library)

Prerequisites: None.

Table of contents:

- Introduction to Law and Legal Studies
- Introduction to the EU Law
- Introduction to the EU GDPR
- The concept of data; personal, sentitive and economic data; big data
- Property of data, choices in the management of data
- The right to be forgotten
- Civil and criminal aspects of profiling activity
- Automatic data processing, human responsabilities
- The Data Protection Officer and DP Authorities
- Civil and criminal protection of privacy
- Sanctioning powers and system
- Open Data for the public interest
- Big data (collection, analysis, processing) and their influence on fundamental rights
- Digital Surveillance
- Facial Recognition: Open Issues
- Disinformation
- Artificial Intelligence in the EU law

Assessment breakdown:
Written Exam

University: Universitá degli Studi di Padova (UniPd)
Department: Dipartimento di Matematica
Course ID: StochMethod
Course name: Stochastic Method
Name and email address of the instructors: Marco Ferrante (marco.ferrante@unipd.it)
Web page of the course: https://en.didattica.unipd.it/off/2025/LM/SC/SC2738/003PD/SCP7079197/N0
Semester: 1
Number of ECTS: 6

Course breakdown and hours:

- Lectures, Exercises, and Laboratory Sessions: 48h.

Goals:
The aim of this course in to introduce tools from the theory of Probability and Stochastic Processes that have high impact in the study of networks as well as algorithmic and computational tools.

Learning outcomes:
The students will acquire the skills to model stochastic processes that have high impact in the study of networks, as well as they will learn how to represent these processes in software tools and derive relevant aspects.

Readings and text books:
Paolo Dai Pra, Stochastic Methods for Data Science, Lecture notes, 2016.
Michele Pavon, Discrete probabilistic models. Lecture notes, 2015.

Prerequisites: Basic notions of differential and integral calculus, linear algebra and probability.

Table of contents:

1. **Probability Reviews**
   - Probability spaces, sigma-field, properties of probability
   - Conditional probability and independence
   - Random variables, expectation and conditional expectation
   - Law of Large Numbers and Chernoff bounds
   - Approximation of probability distributions
2. **Discrete-Time Markov Chains**
   - Discrete-time Markov chain: definition and classification of the states
   - Stationary distribution and convergence results
3. **Discrete-Time Martingales**
   - Properties of conditional expectation
   - Definition of martingale, submartingale, and supermartingale
   - Stopping times

Assessment breakdown:
Written Exam

Chapter 5

TU Wien — Knowledge-Driven Data Management and Intelligence (3rd semester)

This semester focuses on the specialisation in two critical components of "Knowledge-Driven Data Management and Intelligence" namely data management and artificial intelligence. On the side of data management, we deepen the knowledge in the critical area of graph data management. On the side of AI, we provide a broad preparation covering both Symbolic AI and Subsymbolic AI. The semester is composed of the following courses.

Students take the mandatory courses GraphData, GraphAI, DeclAI, and choose 12 ECTS among NLP, AIE, RDM, TEP, and TEC for a total of 30 ECTS.

5.1 Detailed Description of Curriculum

University: TU Wien
Department: Institute of Logic and Computation
Course ID: GraphData
Course name: Management of Graph Data
Name and email address of the instructors: Katja Hose (katja.hose@tuwien.ac.at)
Semester: 3
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 14 hours
- Assignments/Project: 119 hours
- Exam preparation: 15 hours
- Exam: 2 hours

Goals:
This course prepares students for practical challenges in managing, querying, and integrating graph data in complex interconnected systems, focusing on the latest technologies and frameworks in this rapidly evolving area. Students explore RDF-based knowledge graphs along with schema definitions using RDF-Schema and OWL, data quality standards such as SHACL, advanced querying in SPARQL, and provenance management and tracking. The students will also work with the property graph model covering basic schema concepts and powerful query languages such as Cypher and PGQL. The course further addresses practical aspects of applying and working with knowledge graph, e.g., semantic interoperability, knowledge-graph-based data integration, schema and query conversion, unified graph data models, evolving knowledge graphs, and real-world applications, demonstrating the transformative potential of semantic data technologies.

Learning outcomes:
After successful completion of the course, students are able to

- explain key graph data models and query languages
- formulate complex queries using graph query languages
- explain principles of query optimization
- apply schema languages tailored for graph data
- ensure data quality through shape constraints
- use state-of-the-art graph data technologies in popular applications and use cases

Readings and text books:
No specific textbook will be used, but reference to academic reading will be given in-context.

Prerequisites:
None, apart those of the program itself.

Table of contents:

- Key concepts of graph data models and schema languages
- Query formulation, execution, and optimization
- Ensuring graph data quality through validation and shape constraints
- Methods for tracking provenance and ensuring traceability of graph data
- Recent advances and trends in graph data technologies

Assessment breakdown:
Individual assignments/project, exam

University: TU Wien
Department: Institute of Logic and Computation
Course ID: GraphAI
Course name: Graph AI
Alternative name at TUW: Knowledge Graphs
Name and email address of the instructors: Emanuel Sallinger (sallinger@dbai.tuwien.ac.at)
Semester: 3
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 30h
- Exercises and Project: 120h

Goals:

The aim of this is course is to gain a deep understanding of Graph-based AI and the connection of Knowledge Graphs to Artificial Intelligence, Machine Learning, Deep Learning and Data Science. Particular covered items are Knowledge Graph Embeddings, Graph Neural Networks and the combinations of LLMs and Knowledge Graphs. It is divided into three blocks:

- Data and AI Representations of (Knowledge) Graphs
  logic- and ML-based: embeddings, logical knowledge, graph neural networks,
- Systems for Knowledge Graphs
  scalablility and reasoning: reasoning in knowledge graphs
- Applications of Knowledge Graphs
  real-world enterprise AI

An overarching aim of the course is to understand the connections in Graph-based AI, namely between Knowledge Graphs (KGs), Artificial Intelligence (AI), Machine Learning (ML), Deep Learning and Data Science. A particular focus is of gaining a broad understanding of all of the following learning outcomes, coming from throughout the database, semantic web, machine learning and data science communities.

Learning outcomes:

Learning outcomes (LOs) are structured into exactly the above three blocks. A particular focus is of gaining a broad understanding of all of the following learning outcomes, coming from throughout the database, semantic web, machine learning and data science communities.

Representations

The aim in this part is to understand and apply the predominant representations of knowledge and data in Knowledge Graphs.

- (LO1) Describe and apply Knowledge Graph Embeddings
- (LO2) Describe and apply logical knowledge in KGs
- (LO3) Describe and apply Graph Neural Networks and Graph Transformers
- (LO4) Compare different Knowledge Graph data models from the database, semantic web, machine learning and data science communities.

This, crucially, includes the connections between using these types of knowledge in one Knowledge Graph.

Systems

The aim in this part is to be able to design and apply systems that manage Knowledge Graphs.

- (LO5) Design and implement architectures of a Knowledge Graph
- (LO6) Describe and apply scalable reasoning methods in Knowledge Graphs
- (LO7) Apply a system to create a Knowledge Graph
- (LO8) Apply a system to evolve a Knowledge Graph

The latter two learning outcomes (together with LO11) provide a typical life-cycle of Knowledge Graphs: getting data into a KG, i.e., creating it (LO6), evolving a KG into a new one (LO7) and getting data out of a KG by providing services based on it (L11). Note that the term “life-cycle” is used loosely here, as in many Knowledge Graphs, providing applications is not necessarily the end of the life-cycle, but part of an on-going activity.

Applications

The aim of this part is to understand and design applications of Knowledge Graphs.

- (LO9) Describe and design real-world applications of Knowledge Graphs
- (LO10) Describe financial Knowledge Graph applications
- (LO11) Apply a system to provide services through a Knowledge Graph
- (LO12) Describe the connections between Knowledge Graphs (KGs), Machine Learning (ML) and Artificial Intelligence (AI)

Readings and text books:
No specific textbooks will be used, but reference to academic reading will be given in-context.

Prerequisites:
None, apart those of the program itself.

Table of contents:
Following the didactic principles of constructive alignment, content, learning outcomes and assessment are aligned, hence see the learning outcomes.

Assessment breakdown:

Summative assessment is by one item: the project portfolio. In particular, it:

- gives a clear, transparent basis for the achieved marks, and puts control of demonstrating the achieved learning outcomes into one's own hands.
- allows diversity in how to demonstrate the learning outcomes: while a typical way to demonstrate that would be a practical project portfolio, learners who are pursuing a theoretical direction will be able to demonstrate that through a theoretical project portfolio.
- allows for self-regulated learning: while the portfolio is the single assessment item, it is the witness of a longer learning process, and can be created - using the principle of "patchwork" text - throughout the learning process, or in one block at the end, depending on the learning type.

University: TU Wien
Department: Institute of Logic and Computation
Course ID: SymbAI
Course name: Declarative Knowledge and Symbolic AI
Alternative name at TUW: Processing of Declarative Knowledge
Name and email address of the instructors: Magdalena Ortiz (magdalena.ortiz@tuwien.ac.at), Mantas Šimkus (mantas.simkus@tuwien.ac.at) , Thomas Eiter (thomas.eiter@tuwien.ac.at)
Semester: 3
Number of ECTS: 6

Course breakdown and hours:

- 36 hours in-person lectures,
- 72 hours of exercises,
- 42 hours project or exam.

Goals:
This course introduces methods and techniques for representing and declaratively processing knowledge. The representation takes the form of sentences in a symbolic language whilst the processing is done by procedures accessing these sentences. The students learn how problem specifications can be expressed in logic-based representation languages, while specialised reasoning engines can be used to process these sentences. In contrast to procedural approaches, which require programmers to explicitly define how computations should be performed, declarative methods allow reasoning systems to determine the necessary computations based on the input specification. By separating problem specifications and domain knowledge from implementation details, declarative methods are a powerful alternative offering benefits such as improved solution quality, reusability, transparency, and explainability.

Learning outcomes:
After a successful completion of the module, students will be able to

- specify declaratively (possibly recursive) Datalog queries and ASP programs,
- write ASP programs and classify them by their language features and computational properties,
- understand the principles of Datalog query optimisation via magic set transformation,
- use description logics (DLs) for writing ontologies, as well as distinguish and compare different DLs,
- name differences between (logic programming) rules and ontologies,
- model problems in hybrid languages that combine rules and ontologies,
- differentiate between declarative and procedural problem-solving approaches,
- extract accurate and complete problem specifications from natural language descriptions,
- develop and optimise solutions for different types of declarative specifications, and
- critically assess the suitability of different declarative solving techniques for different real-world problems.

Readings and text books:
No specific textbooks will be used. Academic reading materials will be provided during the course.

Prerequisites: *None, apart those of the program itself.*

Table of contents:
In this module, we consider different formalisms for representing and processing declarative knowledge, with particular emphasis on

- Principles of declarative knowledge processing
- Datalog and extensions
- Optimization of Datalog queries
- Answer-set programming
- Modeling problems in ASP
- Description logics as languages for ontologies
- Computational aspects of reasoning
- Hybrid languages combining rules and ontologies
- Formalisms for contextual reasoning

Assessment breakdown: Assignments, project, and final exam. The detailed modalities and criteria are described at the beginning of the semester.

University: TU Wien
Department: Institute of Logic and Computation
Course ID: AIE
Course name: AI Ethics
Name and email address of the instructors: Kees van Berkel (kees.van.berkel@tuwien.ac.at)
Semester: 3
Number of ECTS: 6

Course breakdown and hours:

- Lectures and seminars 36h
- Preparation for the seminars 36h
- Engagement 10h
- Preparation for the presentation 10h
- Writing reviews for and processing reviews of other students 8h
- Final paper 50h

Goals:
With the rapid developments in Artificial Intelligence (AI), we find an inevitable increase in AI's impact on our lives, posing various ethical challenges for developers, policy-makers, and society at large. This has led to a general awareness that AI systems must be aligned with human values, ethics, and laws. This module introduces various ways in which ethics and AI meet. It enables students to develop basic skills for identifying and critically reflecting on AI-specific ethical challenges, and for incorporating critical thinking on these topics in their work and further studies.

This module deals with the various ethical challenges and topics arising in the context of AI. Besides an in-depth introduction to ethical frameworks and AI ethics themes, the module focuses on the development of critical reflection and argumentation/ discussion skills.

Learning outcomes:
After successful completion of the course, students are able to

- explain and summarize central concepts and problems in AI ethics (e.g., AI alignment, algorithmic bias);
- describe fundamental ethical theories (e.g., utilitarianism, care ethics) and illustrate their application to AI ethics topics through examples;
- analyze and construct a critical ethical reflection on central technical and philosophical challenges in AI ethics (e.g., fairness, sustainability, AI alignment);
- formulate and defend clearly structured, state of the art informed arguments on key AI ethics questions (e.g., the right to explanation, data privacy), engaging with counterarguments; and
- justify the importance of ethical reflection in AI development, using ethical reflection and/or case studies to support claims.

Students will acquire critical thinking skills that are strongly rooted in humanity studies. They will also learn how to study text on policy, law, and regulations concerning AI.

Readings and text books:
No lecture notes are available

Prerequisites: None

Table of contents:

- Novel ethical challenges in AI such as disinformation, explainability, sustainability, data-privacy, autonomy
- Methods such as embedded ethics, multi-scale ethics, AI ethics audits, AI alignment, human-in-the-loop AI
- Ethical theories including consequentialism, utilitarianism, virtue ethics, deontology, care ethics, and critical methods such as feminist ethics

Assessment breakdown:
Final essay (60%), Presentation (20%), Reviewing (10%), Forum participation (10%)

University: TU Wien
Department: Institute of Information Systems Engineering
Course ID: NLP
Course name: Natural Language Processing and Information Extraction
Name and email address of the instructors: Allan Hanbury
Semester: 3
Number of ECTS: 3

Course breakdown and hours:

- Lectures: 24h
- Exercises and Project: 51h

Goals:
Gain understanding of the topics:
- Basics of text processing: segmentation, tokenization, decompounding, stemming, lemmatization; regular expressions
- N-gram language modeling, simple classification tasks in NLP
- Part-of-speech tagging, named entity recognition, and shallow parsing with Hidden Markov Models
- Syntactic representations and syntactic parsing
- Basics of natural language semantics
- Neural network basics. Feed forward networks and recurrent neural networks
- Sequence modeling and sequence-to-sequence models.
- Neural language modeling. Word vectors and contextualized language models.
- Information extraction tasks: entity recognition, relation extraction, knowledge base population
- Information extraction applications: summarization, question answering, chatbots

Learning outcomes:
After successful completion of the course, students are able to extract structure from natural language data by applying standard methods for text segmentation, word and sequence tagging, or syntactic parsing. They will have a high-level overview of the most important rule-based and learning-based approaches to each task and the standard methods for evaluating them. Students will gain a fundamental understanding of artificial neural networks and methods for training them, with a special emphasis on architectures for processing sequential data, allowing them to solve a variety of NLP tasks with deep learning. An overview of information extraction tasks will be given, allowing students to approach various problems involving the extraction of structured information from unstructured text data. A survey of common specialized IE tasks is also provided, acquainting the students with some of the most common NLP applications.

Readings and text books:
Announced in the course.

Prerequisites:
None, apart those of the program itself.

Table of contents:
As per the goals.

Assessment breakdown:
15% for Milestone 1, 15% for Milestone 2, 50% for the final solution, 10% for the presentation, 10% for the management summary as described below

Term project (done in groups) with milestones. The performance assessment consists of three partial achievements: Milestone 1, Milestone 2, Final submission

The three partial achievements are compulsory. As each of these three partial achievements listed above is a partial achievement that must be completed over a longer period of time (at least two weeks are planned for each exercise), there is no possibility of repetition for this partial achievement.

University: TU Wien
Department: Institute of Information Systems Engineering
Course ID: RDM
Course name: Introduction to Research Data Management
Name and email address of the instructors: Tomasz Miksa (tomasz.miksa@tuwien.ac.at)
Semester: 3
Number of ECTS: 3

Course breakdown and hours:

- Lectures: *integrated*
- Exercises and Project: *integrated*

Goals:
Gain understanding of the topics:

- Reproducible Research
- Metadata & Data Packaging
- Machine-actionable Data Management Plans
- Research Data Management Services
- Repository Systems
- Data Identification and Citation
- Digital Preservation Challenges, Preservation Actions
- Preservation Planning & OAIS Standard
- FAIR principles
- Legal & Ethical Aspects of Research Data Management

Learning outcomes:
After successful completion of the course, students are able to:

- explain the research & data lifecycles
- review existing Data Management (DM) practices
- create a Data Management Plan (DMP)
- differentiate between different software and data license types
- explain what persistent identifiers are (e.g. DOI)
- apply FAIR principles in practice
- explain how findability and interoperability of data can be achieved
- understand the function of metadata and standards for efficient (research) data management
- describe components of a Research Data Management (RDM) infrastructure
- discuss differences between repository systems
- describe the scope of RDM policies and how they drive DM activities and obligations
- explain challenges in digital preservation
- plan, assess, and implement reproducible experiments
- understand and apply Open Science principles

Readings and text books:
Will be announced in the course.

Prerequisites:
None, apart those of the program itself.

Table of contents:
As per the goals.

Assessment breakdown:
A set of assignments must be solved to pass this course. Assignments may include solving challenges in the management of data applying the recommendations and guidelines from the lecture, using domain specific tools and services for data management, and a subsequent cap presentation.

University: TU Wien
Department: Zentrum für strategische Lehrentwicklung
Course ID: TEP
Course name: Technical English Communication
Name and email address of the instructors: *Announced each term.*
Semester: 3
Number of ECTS: 3

Course breakdown and hours:

- Lectures: *integrated*
- Exercises and Project: *integrated*

Goals:
This course prepares students for the transition from academic to professional life. You'll learn how to interview and prepare an application effectively by learning the intricacies of a compelling resume and cover letter. Through flipchart presentations and critical thinking workshops, this course will expand your technical, scientific and business vocabulary. In addition to avoiding typical Germanisms, there is a special focus on reading scientific articles, writing essays, grammar, pronunciation and punctuation.

Workshops are held on writing resumes, cover letters and interviewing. Students submit their applications and each student has a 5 minute personal interview. Poster presentations are held in pairs to differentiate between technical and scientific vocabulary. There is a written and an oral intermediate and final exam with a focus on grammar, pronunciation, choice of words and writing short essays.

Learning outcomes:
After successfully completing the course, students are able to speak and write about technical issues in English. Through group work, debates, poster presentations, and exams on a variety of technical and business topics, you can expand your technical vocabulary, learn to use technical and academic English naturally and confidently, and improve your grammar and pronunciation.

Readings and text books:
Announced in the course.

Prerequisites:
None.

Table of contents:
Workshops are held on CV and cover-letter writing, as well as interviewing. Students submit their job applications, and every student has a 5-minute personal interview. Poster presentations on the differentiation of technical and scientific vocabulary are given in student teams. There is a written and oral midterm and final exam, which focuses on grammar, pronunciation, word choice, and short-essay writing.

Assessment breakdown:
Out of a total of 400 points:
(50 points) CV and Cover Letter
(50 points) Team Introduction and Word-Differentiation
(100 points) Midterm Exam
(50 points) 2-minute Job Interview
(50 points) Team Debate, Project Proposal, or Critical Thinking Poster-Talk
(100 points) Final Exam

University: TU Wien
Department: Zentrum für strategische Lehrentwicklung
Course ID:TEP
Course name: Technical English Presentation
Name and email address of the instructors: *Announced each term.*
Semester: 3
Number of ECTS: 3

Course breakdown and hours:

- Lectures: *integrated*
- Exercises and Project: *integrated*

Goals:
This course specializes in presentations. Each student gives a 10-minute technical presentation with a question session and feedback session. Students gain practice with various presentation media: flipchart, blackboard and powerpoint. Particular attention is paid to the following topics: Presentation structure and message: take-home message, dealing with "stage fright", creating graphics effectively, giving feedback and thereby learning the intricacies of presenting, building technical vocabulary, correct pronunciation of technical words, Spelling & distinction of terms.

Each student introduces themselves in 1 to 2 minutes. Combined with an English assessment quiz, this gives an overview of the student's presentation and English skills. Each student then gives a ten-minute presentation on a technical, scientific or business topic of their choice. All students learn to give effective feedback on the following topics: Structure - Story & Message, Visuals, Body Language, Speech & Articulation. There is a written and oral intermediate exam testing grammar, pronunciation, and the principles of effective presentation. Each student then gives a 1-2 minute elevator pitch, followed by the final exam.

Learning outcomes:
After successfully completing the course, students are able to present themselves and their work effectively in English. The students are coached in speaking technique and pronunciation, body language and stage presence. The clear communication of scientific research is in the foreground. After this course, students will be able to perform and present more fluently and confidently in English.

Readings and text books:
Announced in the course.

Prerequisites:
None.

Table of contents: Every student gives a 1-to-2 minute Self-Introduction, which in combination with an English assessment quiz provides a baseline of the student's abilities in both presentation and English. Each student then gives a 10-minute presentation on a technical, scientific, or business topic of their choice. All students train the art of giving corrective feedback on the topics: Structure - Story and Message, Visuals, Body Language, and Speech and Articulation. There is a written and oral midterm exam, where grammar, pronunciation, and the principles of effective presentation are tested. This is followed by each student giving a 1-2 minute Elevator Pitch, followed by the final exam.

Assessment breakdown:
Out of a total of 400 points:
(20 points) 1-2 minute Self-presentation
(100 points) 10-minute Presentation on a Technical or Scientific Topic
(15 points) Feedback: Structure, Story, and Message
(15 points) Feedback: Visuals
(15 points) Feedback: Body Language
(15 points) Feedback: Speech & Articulation
(20 points) 1 Role: Chairperson, Camera Technician, Heckler, or Coordinator
(100 points) Midterm Exam: written and oral (Pronunciation)
(100 points) Final Exam: written and oral (Elevator pitch)

Chapter 6

Lyon 1 — Data and Intelligence for Heterogeneous Systems (3rd semester)

6.1 Motivation and Background

Smart systems are intelligent, data-driven systems designed to perform complex processing and control tasks by collecting, processing, and analyzing vast amounts of data. This data is highly heterogeneous, encompassing structured and unstructured forms such as text, images, and processes. To enable meaningful analysis and prediction, the data must undergo preparation, processing, and modeling to support advanced analytical and inference techniques used by smart systems.

These systems face significant computational challenges, including the storage and processing of streaming data from sensors and IoT devices, as well as applying symbolic and statistical learning to these datasets. Additionally, the development of services and models that leverage this data plays a crucial role in enabling the intelligent functionality of smart systems.

The core of the master's training program is centered around the seamless integration and processing of diverse data types—from structured data to graph-based models and images, including both static and dynamic datasets. The program also emphasizes advanced inference techniques and artificial intelligence (AI) processes that drive analysis, prediction, and forecasting capabilities in smart systems.

Data-driven intelligent systems form the backbone of Industry 4.0 and Digital Transformation, with widespread applications across various sectors such as healthcare, education, energy efficiency, transportation, and climate change.

The master's program has two primary objectives:

- To prepare students for PhD studies and research careers in both academia and industry.
- To equip students with strong research methodologies, while deepening their expertise in interdisciplinary areas of computer science, including data management, graphics and visual computing, artificial intelligence, service-oriented computing, natural language processing, and human-centered technologies.

6.2 Proposal of Specialization

This program equips students with a deep understanding of advanced data science techniques, focusing on real-time data processing, machine learning, graph data management, natural language processing (NLP), and human-centered approaches. In addition to technical knowledge, students will develop essential research and professional skills through soft skills training and research methodology courses.

The master's program provides a comprehensive and well-rounded curriculum, equipping students with a diverse range of competences and skills essential for modern data science and research. The curriculum includes:

- **In-depth knowledge of Big Data processing systems**, focusing on managing heterogeneous data and real-time data analytics. Students learn how to handle large-scale data processing challenges with cutting-edge tools and frameworks like Apache Kafka, Flink, and Spark.
- **Mastery of machine learning techniques and applications**, with a thorough exploration of both supervised and unsupervised learning methods. The course covers classic algorithms like linear regression and decision trees, as well as advanced techniques such as XGBoost and neural networks.
- **Advanced study of Natural Language Processing (NLP)**, emphasizing its practical applications in data management. Students gain hands-on experience with modern NLP libraries and techniques, such as transformers, BERT, and large language models, for tasks like text classification and sentiment analysis.

- **Incorporating human-centered approaches in AI applications**, exploring how to integrate the human perspective into the design and development of systems. This includes working with human-computation models, crowdsourcing, and human-in-the-loop algorithms.
- **Advanced graph data processing**, covering both the systems and languages necessary for managing and querying complex graph structures. Students learn to use graph query languages like Cypher and engage with real-world graph databases such as Neo4j to solve practical challenges.
- **Research methodology and critical thinking**, where students develop the capability to formulate research questions, conduct literature reviews, and design solutions to complex research problems. This skillset is crucial for academic or industry-focused research.
- **Introduction to the European research grant framework**, focusing on understanding the structure of programs like Horizon Europe and Marie Curie grants, with guidance on writing proposals and managing research projects.

6.2.1 Specialization Curriculum

The curriculum at Lyon 1 focuses on advanced topics in data science, equipping students with essential skills in real-time data processing, machine learning, graph processing, and advanced natural language processing, alongside important soft skills for research and professional development.

- **Real-Time Data Processing and Analytics** delves into big data management, emphasizing low-latency and high-throughput systems like Apache Flink, Kafka, and Spark. Students learn to handle streaming data, continuous queries, and real-time graph data management.
- **Machine Learning Techniques and Applications** covers both theory and practical applications of machine learning, including supervised and unsupervised methods, deep learning, and graph data mining. The course includes hands-on projects using Python to apply techniques to real-world datasets.
- **Natural Language Processing** introduces NLP fundamentals, focusing on both classical and deep learning techniques. Students gain practical experience with libraries such as NLTK, spaCy, and Hugging Face, and explore advanced topics like transformers, BERT, and large language models.
- **Big Graph Processing Systems** focuses on advanced graph data models, querying, and optimization techniques. Students learn to work with property graphs and RDF data, using open-source tools like Neo4j for real-world data projects.
- **Human-Centered AI** teaches how to integrate human perspectives into data-intensive applications, focusing on human-computation systems, crowdsourcing, and human-in-the-loop algorithms.
- **Research Methodologies (1 and 2)** courses focus on developing research capabilities, including how to formulate research questions, conduct literature reviews, and design experiments. There is also training on writing research grants and managing research projects.

6.2.2 Academic Calendar and Marking System

The schedule for the specialization is organized as follows:

- **Start of the semester:** Second week of September
- **End of the semester:** Second week of February
- **Examination period:** Last two weeks of February
- **Master's project period:** March to August

Throughout the semester, course instructors have the flexibility to schedule midterm assessments. During the official examination period, each course must provide two examination opportunities.

Exams are graded on a scale of 0 to 20, with a minimum passing grade of 10 required for successful completion.

6.2.3 University Fees

The annual enrollment fees of University of Lyon1 is € 250, and they cover the entire semester.

6.2.4 Master Thesis

Students will complete their final Master's project at Lyon 1 under the guidance of a professor or lecturer from the master's program. The project can be developed in collaboration with a company, focusing on addressing specific questions or challenges posed by the company. During this process, students will apply the knowledge and skills gained throughout the first and second years of the program.

6.2.5 Language Courses

To help the students improve their French skills, the University Service for Language Teaching (Service Commun d'Enseignement des Langues or "SCEL") at Lyon1 offers free French as a Foreign Language (Français Langue Etrangère or "FLE") courses for beginners to advanced French language learners.

6.2.6 Student Support and Accommodation

Lyon 1 offers several services to international exchange students, including

- Meal: On-campus canteens with affordable meal deal.
- Housing: Exchange students have the possibility to request a spot in public university housing (fr. residence CROUS).
- Cultural Activities: Lyon 1's Astrée theatre, located on the Lyon Tech la Doua Campus, is home to a company combining the arts and sciences. Access to the Astrée Theatre is free for Lyon 1 students.
- Sport. The University Physical and Sports Activities Service (fr. Service Universitaire des Activités Physiques et Sprotives, or "SUAPS") coordinates and manages the university's student sports activities. The SUAPS offers teaching in some thirty disciplines, from initiation to improvement, and also provides the opportunity to train and take part, within the framework of the Lyon1 Sports Association, in competitions organized by the French University Sports Federation (FFSU).

6.3 Key People

- **Angela Bonifati** is the director this master program at Lyon 1. She is a Distinguished Professor of Computer Science at Lyon 1 University and the CNRS Liris research lab, where she leads the Database Group. She is also an Adjunct Professor at the University of Waterloo in Canada since 2020 and a Senior member of the French University Institute (IUF) from 2023. Her research interests include data management, graph databases, knowledge graphs, data integration, and their applications in data science and machine learning. She has received the IEEE TCDE Impact Award 2023 and the ACM Research Highlights Award 2023. Angela is the Program Director of the International Master in Computer Science (DISS) at UCBL since 2022.
- **Andrea Mauri** is a Junior Professor (Chaire de Professor Junior) at Lyon 1 University and the CNRS Liris research lab. His research focuses on the intersection of Human-Computer Interaction and Data Management. He investigates how to integrate human perspectives and computational methods in designing, developing, and deploying data-intensive applications.
- **Rémy Cazabet** is an Associate Professor at Lyon 1 University and the CNRS LIRIS lab in the Data Mining and Machine Learning team. His research applies data analysis and AI to network data, including temporal and spatial aspects. He collaborates with researchers from other fields such as economics and epidemiology, and partners with companies for multiple projects. He teaches Machine Learning and Data Mining in several masters' programs in Lyon, including the DISS Master.
- **Dana-Maria Daia** has been working in International Relations since 2006 and has extensive experience in managing international programs. As Head of the European Programmes Office, she coordinates Lyon 1 University's participation in European academic and mobility projects, focusing on institutional cooperation and internationalization strategy. She works closely with the Vice-Rectors for International Relations and is the Institutional Coordinator for the Arqus European University Alliance.
- **Marcelly de Sousa Magalhaes** is an Erasmus+ project officer at UCBL. She is skilled in the administrative and financial management of European and international projects financed by European funds. Marcelly holds a Master's degree in European Studies with a specialization in law and a Bachelor's degree in International Relations from the University of Geneva. She supports the administrative and financial activities of the project at UCBL.

6.4 Detailed Description of Curriculum

University: Université Claude Bernard Lyon 1 (Lyon1)
Department: Departement d'Informatique
Course ID: RTProc
Course name: Real-time Data Processing and Analytics
Name and email address of the instructors: Angela Bonifati (angela.bonifati@univ-lyon.fr)
Web page of the course: http://master-info.univ-lyon1.fr/DISS/#ues
Semester: 1
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 36h.
- Exercises and Project: 24h.

Goals:

This course focuses on the real-time aspects concerning big data management solutions. In particular, the students will learn how modern processing techniques address low latency and high-throughput velocity requirements. On the OLAP side, this course studies compare and critically applies solutions for streaming data management, e.g., Apache Flink, Apache Kafka, and Spark. On the OLTP side, the course will discuss modern database techniques, called NewSQL, to reduce transaction latency without neglecting strong consistency guarantees. On the practical side, the course discusses techniques for query processing in real-time (continuous queries and complex event processing), the role of SQL and declarative languages, and data preparation techniques and algorithms to handle complex, massive datasets as input to subsequent data intelligence tasks. Finally, the course will present open research direction for extending real-time data management to complex data like Streaming graphs and evolving knowledge graphs.

Learning outcomes:

During the course, the students will be able to use and experiment with existing data processing systems and get familiar with the latest research advances on this topic. At the end, the students acquire skills related to:

1. Real-time big data analytics with modern tools, e.g., Apache Flink;
2. continuous querying and complex event processing using declarative languages;
3. the challenges and opportunities of managing complex data like graphs in real-time.

Readings and text books:

- Kleppmann, Martin. Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems. " O'Reilly Media, Inc.", 2017.
- Stopford, Ben. Designing event-driven systems. O'Reilly Media, Incorporated, 2018.
- Akidau, Tyler, Slava Chernyak, and Reuven Lax. Streaming systems: the what, where, when, and how of large-scale data processing. " O'Reilly Media, Inc.", 2018.

Prerequisites:

1. Experience with SQL and Java Programming
2. Big Data System Deployment: Docker, Kubernetes, Yarn

Table of contents:

- Stream Processing Fundamentals:
  - Ingestion (Apache Kafka)
  - Micro-Batching vs Continuous Processing
  - Scaling real-time systems
  - Handling Stream Imperfections: Data Quality and Out of Order Handling
- Continuous Querying
  - Declarative Languages: SQL and More
  - Continuous Semantics: Windowing vs Complex Event Processing
  - NewSQL vs Streaming Databases: VoltDB vs RisiwingWave
- Advanced Stream Processing
  - Processing Streaming Graph
  - Declarative Languages: Beyond SQL

Assessment breakdown:

The student examinations will be carried out through a project that requires satisfying a continuous information need in real-time using either of the frameworks presented in class. The students are required to present at least two different design and present a critical comparison of their design choice based on what was learnt in class.

University: Université Claude Bernard Lyon 1 (Lyon1)
Department: Departement d'Informatique
Course ID: MLTA
Course name: Machine Learning Techniques and Applications
Name and email address of the instructors: Rémy Cazabet (remy.cazabet@univ-lyon1.fr)
Web page of the course: http://cazabetremy.fr/Teaching/DISS/ML.html
Semester: 1
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 36h.
- Exercises and Project: 24h.

Goals:
The class offers an overview of machine learning and statistical thinking, with a final in-depth study on a recent research question. It tackles both theory and practice, with half of the hours devoted to projects and Python practicals. It presents the latest machine learning methods, spanning from classical supervised and unsupervised methods to deep neural networks and large language models, along with graph data mining and temporal aspects.

Learning outcomes:

1. Initiation to Data cleaning & manipulation, data exploratory analysis and statistical description
2. Knowledge of the main classic supervised ML methods, from linear regression to XGBoost
3. Knowledge of the main unsupervised clustering methods, including DBSCAN and Gaussian Mixture Models.
4. Data Mining in Graphs (community detection, centrality scores, etc.)
5. Deep Learning
   (a) Introduction: perceptron, gradient descent, convolutions, hyperparameters, etc.
   (b) Focus on LLMs (RNN, generative AI, Attention/Transformers, etc.
6. Focus on streaming and temporal methods

Readings and text books:

- Burkov, A. (2019). The hundred-page machine learning book (Vol. 1, p. 32). Quebec City, QC, Canada: Andriy Burkov.
- Aggarwal, C. C. (2015). Data mining: the textbook (Vol. 1). New York: springer.
- Menczer, F., Fortunato, S., & Davis, C. A. (2020). A first course in network science. Cambridge University Press.
- Labonne, M., (2023) Hands-On Graph Neural Networks Using Python. Packt Publishing.

Prerequisites: Knowledge of programming languages, preferably Python. Basic probability theory, linear algebra.

Table of contents:

- Supervised classic: linear regression, logistic regression, decision tree, random forest, AdaBoost, XGBoost. Evaluation for regression and classification.
- Unsupervised classic: k-means, Gaussian Mixtures, DBSCAN. Evaluation: elbow method, Silhouette score.
- Deep Neural Networks: gradient descent, back-propagation, CNN, Autoencoders, RNN, Transformers, LLMs, Graph Neural Networks.
- Graph Data mining: Community detection, Centralities.
- Dynamic, Temporal & Streaming: dynamic clustering, time series forecasting.

Assessment breakdown:
Written exam (50%), Literature review (10%) and a project (40%) requiring applying one or several aspects of the class to a chosen dataset.

University: Université Claude Bernard Lyon 1 (Lyon1)
Department: Departement d'Informatique
Course ID: Research1
Course name: Research Methodologies 1
Name and email address of the instructors: Andrea Mauri (andrea.mauri@univ-lyon1.fr)
Web page of the course: http://master-info.univ-lyon1.fr/DISS/#ues
Semester: 1
Number of ECTS: 6

Course breakdown and hours:

- Lectures: 36h.
- Exercises and Project: 24h.

Goals:
This course will let students become familiar with the research methodology. It will also let them acquire the necessary skills to conduct a successful investigation of the related work, experiment design, and the writing of a research problem statement, prepare the outline of a scientific paper and present their research.

Learning outcomes:
At the end of the course students will acquire skills related to:

1. How to identify a problem worth studying and how to define a problem statement.
2. How to write a proper research question.
3. How to navigate the state-of-the-art and write a literature review.
4. Identify which kind of research contribution fits a specific problem.
5. Design an experiment.
6. Perform both quantitative and qualitative data analysis.
7. Present their results.
8. Identify ethical issues in research and how to conduct ethical research.

Readings and text books:

- Alvesson, Mats; Sandberg, Jorgen. 2013. Constructing Research Questions: Doing Interesting Research. 1st ed. SAGE Publications Ltd.
- Lazar, Jonathan; Feng, Jinjuan Heidi; Hochheiser, Harry. 2010. Research Methods in Human-Computer Interaction. Wiley Publishing.
- Strunk, William I. 1999. The Elements of Style. 4th ed. Upper Saddle River, NJ: Pearson.
- Research articles shared during class.

Prerequisites: Prior knowledge of statistics and programming can be useful for the part of the course dealing with quantitative data analysis. Given the essay-heavy nature of the course, good proficiency in written English is a plus.

Table of contents:

- What is research? Basic vs Applied research. The scientific method.
- The different kinds of research contribution. Empirical, methodological, theoretical, dataset, artifact, survey, and opinion.
- The different types of research questions. Descriptive, comparative, explanatory, and normative.
- How to find and write a good research question.
- How to write a problem statement. The Macro-Meso-Micro framework.
- How to write a literature review, using support tools such as Zotero and Connected Papers.
- The element of experiment design. Dependent, independent, control, and confounding variables. The null hypothesis. Reproducibility vs replicability. Internal vs external (and ecological) validity.
- Experiment with human subjects.
- Responsible and Ethical Research.
  IRB, informed consent, and data handling.
  Nuremberg Code, EU AI Act, Belmont Report principles.
- Statistical methods for quantitative data analysis. Parametric and non-parametric statistical significance tests.
- Qualitative data analysis. Content analysis, coding, and grounded theory.
- How to publish the data related to your research. The FAIR principles.
- How to write a scientific paper. How to present your results.

Assessment breakdown:
The student examinations will be carried out through a group project that consists to write a small research paper. The students will be evaluated in five different stages of the process: 1) problem statement, 2) literature review, 3) experiment design, 4) data analysis and discussion, and 5) final oral presentation of the project.

University: Université Claude Bernard Lyon 1 (Lyon1)
Department: Departement d'Informatique
Course ID: NLPLLM
Course name: Natural Language Processing and LLMs
Name and email address of the instructors: Andrea Mauri (andrea.mauri@univ-lyon1.fr)
Web page of the course: http://master-info.univ-lyon1.fr/DISS/#ues
Semester: 1
Number of ECTS: 3

Course breakdown and hours:

- Lectures: 15h.
- Exercises and Project: 15h.

Goals:

The goal of the course is to make student understand core NLP concepts and equip students with a solid foundation in Natural Language Processing, including classical and deep learning approaches. The course provides practical experience in developing NLP applications using modern libraries like NLTK, spaCy, and Hugging Face and expose the students to to advanced concepts like Transformers, BERT, and Large Language Models (LLMs).

Learning outcomes:

By the end of this course, students will be able to:

1. Understand the basic concepts of NLP, including text preprocessing, tokenization, and word embeddings.
2. Implement traditional and deep learning techniques for NLP tasks such as text classification, sentiment analysis, and language modeling.
3. Explain and use pre-trained language models like Word2Vec, BERT, and GPT, and apply them to NLP tasks.
4. Fine-tune Large Language Models (LLMs) for specific applications like text generation and summarization.
5. Be aware of the ethical and privacy issue related to (Large) Language Models and AI.
6. Work on a team-based project, using datasets to develop a fully functional NLP system, and present the findings.

Readings and text books:

- Speech and Language Processing*. https://web.stanford.edu/ jurafsky/slp3/ed3book.pdf. Authors: Daniel Jurafsky and James H. Martin. 2000.
- Build a Large Language Model (From Scratch)*. Author: Sebastian Raschka. September 2024.
- Documentation of standard python libraries.
- Research articles shared during class.

Prerequisites:

- Proficiency in Python programming language (or the willingness to pick it up during the course).
- Familiarity with basic ML Algorithms (e.g. Logistic Regression, SVM).
- Basic Understanding of Linear Algebra and Probability (Helpful).

Table of contents:

- Introduction to NLP (text processing, basic linguistic concepts, classical models for NLP)
- Word Embeddings and Feature Representations
- Language Models and Sequence Modeling
- Transformer Architecture and Attention Mechanism
- BERT and Transfer Learning in NLP
- Large Language Models (LLMs) and Generative AI
- Special Topics in NLP

Assessment breakdown:

Individual coding assignment and group project

University: Université Claude Bernard Lyon 1 (Lyon1)
Department: Departement d'Informatique
Course ID: Research2
Course name: Research Methodologies 2
Name and email address of the instructors: Andrea Mauri (andrea.mauri@univ-lyon1.fr)
Web page of the course: http://master-info.univ-lyon1.fr/DISS/#ues
Semester: 1
Number of ECTS: 3

Course breakdown and hours:

- Lectures: 15h.
- Exercises and Project: 15h.

Goals:
The goal of this course is to learn what are the fundamental elements of a research grant, focusing on the Horizon program. In particular the course will cover:

- An overview of the Horizon Program
- Differences between quality, impact and implementation of a research project.
- Deep dive on the implementation and elements of project management.
- Deep dive on how to write a Marie Curie grant.

Learning outcomes:
At the end of the course students will acquire skills related to:

1. Understanding how the Horizon program works.
2. Project management
3. Initial understanding on what are the important element of a Marie Curie grant.

Readings and text books:

- Official material from the European Union.
- Articles and slides shared during class.

Prerequisites: A basic understanding of what research is and concepts explained in the Research Methodologies 1 course.

Table of contents:

- Horizon Europe
- The different pillars
  - ERC, Marie Curie, Research Infrastructure
  - The global challebges
  - Innovative Europe
- Key aspects of a Horizon Grant
- ERC grant
- Marie Curie grant
- Implementation and Project Management

Assessment breakdown:
The student examinations will be carried out through a group project that consists to write a small the implementation part of a Marie Curie project.

University: Université Claude Bernard Lyon 1 (Lyon1)
Department: Departement d'Informatique
Course ID: HumanAI
Course name: Human-Centered AI
Name and email address of the instructors: Andrea Mauri (andrea.mauri@univ-lyon1.fr)
Web page of the course: http://master-info.univ-lyon1.fr/DISS/#ues
Semester: 1
Number of ECTS: 3

Course breakdown and hours:

- Lectures: 15h.
- Exercises and Project: 15h.

Goals:
The key goal of this course is to integrate human perspective and computational methods in the design, development and deployment of data-intensive and AI-based applications. The human-centric perspective within data and AI systems is very important to tackle challenges, such as trustworthiness, explainability and interpretability.

Learning outcomes:
At the end of the course students will acquire skills related to:

1. Identify the requirements for a human-centered applications based on data and AI
2. Design and evaluate a human-centered applications based on data an AI.
3. Determine which human-centered algorithms are appropriate for a specific problem.

Readings and text books:

- Human Computation. Author(s): Edith Law and Luis von Ahn. Synthesis Lectures on Artificial Intelligence and Machine Learning, June 2011, Vol. 5, No. 3.
- A. Marcus and A. Parameswaran. Crowdsourced Data Management: Industry and Academic Perspectives. Foundations and Trends in Databases, vol. 6, no. 1-2, pp. 1–161, 2013.
- Research articles shared during class.

Prerequisites: Basic knowledge of artificial intelligence and/or human computer interaction is advised. Proficiency in at least one programming language.

Table of contents:

- Overview of systems for/with collective intelligence (e.g., recommendation, semiautonomous systems, citizen science, crowdsourcing, and human computation systems)
- Human-in-the-loop vs Human-Centered AI vs Human-AI Collaboration
- How to design algorithms that seamlessly use human input.
- Crowdsourcing and Human Computation (e.g., worker modelling, task modelling, incentives, task assignment, recruitment).
- Fairness in human computation and human-centered AI systems.
- Designing Human-Centered AI systems.

Assessment breakdown:
Group assignment on designing a human-centered application.

University: Université Claude Bernard Lyon 1 (Lyon1)
Department: Departement d'Informatique
Course ID: BigGraph
Course name: Big Graph Processing Systems
Name and email address of the instructors: Angela Bonifati (angela.bonifati@univ-lyon1.fr)
Web page of the course: http://master-info.univ-lyon1.fr/DISS/#ues
Semester: 1
Number of ECTS: 3

Course breakdown and hours:

- Lectures: 15h.
- Exercises and Project: 15h.

Goals:
This module will focus on: advanced graph processing and graph querying for property graphs; standardized graph query languages based on recent ISO/IEC proposals, such as GQL and SQL/PGQ; analytical systems and their languages leading to optimized and cost-based query execution; property graph transformations to enable pre-processing, such as graph data cleaning, graph data integration and advanced processing, such as graph-based causal inference.

Recent advances in complex graph query processing will be presented along with evaluation metrics for efficient graph query processing and analytics.

Learning outcomes:

The students will acquire competences and skills on the following topics:

1. Formalization of property graph graph data models and graph queries
2. Constraint enforcement and data cleaning techniques for property graphs
3. Languages for schema specification for property graphs and RDF data
4. Schema and constraint languages for property graphs
5. Property graph transformations and their applications (e.g. graph-theoretic operations for causal inference)
6. Systems and performance aspects of graph databases

Readings and text books:

- Angela Bonifati, George H. L. Fletcher, Hannes Voigt, Nikolay Yakovets: Querying Graphs. Synthesis Lectures on Data Management, Morgan & Claypool Publishers 2018
- Renzo Angles, Marcelo Arenas, Pablo Barceló, Aidan Hogan, Juan L. Reutter, Domagoj Vrgoc: Foundations of Modern Query Languages for Graph Databases. ACM Comput. Surv. 50(5): 68:1-68:40 (2017)

Prerequisites: Knowledge of relational database languages including query, update and schema/constraint languages (Data Manipulation Languages and Data Definition Languages).

Table of contents:

- Data models for graphs (Property Graphs, PG variants and RDF data models)
- Graph Query languages (e.g. openCypher, stardardized GQL and SPARQL)
- Constraints and transformations for PGs
- Architectures for Big Graph Processing Systems (Scale-out versus Scale-up)
- Schemas and schema discovery for PGs (PG-Schema and SHACL)
- Practical labs on graph queries on real-world data and using open-source graph databases (Neo4j, Gradoop)

Assessment breakdown:
The student assessment will be carried out with (1) multiple choice questionnaires on the theoretical part of the course and (2) a parctical lab project addressing query specification on a real-world dataset using an open-source graph database.